BCS
INFORMATION RETRIEVAL

The Proceedings of the 3rd BCS-IRSG Symposium on Future Directions in Information Access

General Chairs: Massimo Melucci and Ricardo Baeza-Yates
Program Chair: Leif Azzopardi

As part of : ESSIR 2009 padova

Preface

FUTURE DIRECTIONS IN INFORMATION ACCESS

In 2007, the 1st BCS-IRSG Symposium on Future Directions in Information Access was established to provide a forum for early career researchers to present, share and discuss research which is at a more formative or tentative stage. The symposium was run in conjunction with the 6th European Summer School in Information Retrieval (ESSIR) which was held in Glasgow. The second symposium was held in London, UK in September, 2008 and now the symposium is in its 3rd year and again collocated with this year's ESSIR in Padua, Italy.

Symposium Aims

The objectives of the Symposium on the Future Directions in Information Access (FDIA) are:

- To provide an accessible forum for early researchers (particularly PhD students, and researchers new to the field) to share and discuss their research.
- To create and foster formative and tentative research ideas.
- To encourage discussion and debate about new future directions.

Symposium Themes

Why Future Directions? To encourage research that focuses on the potential research lines and paths that can and are being developed, where work presenting the, what if scenarios, possible solutions, pilot studies, conceptual and theoretical work is promoted and discussed.

Why Information Access? To capture the broader ideas of information retrieval, storage and management to include interaction and usage.

FDIA 2009

These proceedings contain the papers presented at the Third BCS IRSG Symposium on Future Directions in Information Access (FDIA2009), held in Padua on the 1st of September 2009 during the 7th European Summer School on Information Retrieval (ESSIR 2009).

This years symposium received 20 paper submissions and 5 poster submissions of which 23 were accepted for publication and presentation. Each paper was reviewed by two senior Information Retrieval researchers who were asked to provide detailed reviews and comments to help steer and guide the research presented. In order to obtain high quality feedback each reviewer was asked to review at most two papers. This year's program included some very interesting and promising lines of research such as the Dialogue Driven IR Systems, Multimodal Interaction in IR, and even Quantum Inspired IR Theory. This was fronted by an inspiring keynote talk by Stefan Mizzaro on a Science 2.0 approach to scholarly publishing and peer review.

The organizers would like to thank: the members of the program committee for all their hard work and effort in providing excellent feedback and reviews, Stefano Mizzaro for delivering a thought provoking key note, all the volunteers of the European Summer School, namely: Marco Bressan, Emanuele Di Buccio, Giorgio Maria Di Nunzio, Marco Dussin, Ivano Masiero, Riccardo Miotto, Nicola Montecchio, Nicola Orio, Michele Scquizzato, Gianmaria Silvello and Francesco Silvestri, as well as, the Administrative Staff at the University of Padua, Maria Bernini, Antonio Camporese, Sabrina Michelotto and Enrico Soncin of the Finance OCE of DEI, and to all the technical staff of DEI for their help and assistance in ensuring the success of this event. Also, we would like to extend our thanks to the BCS-IRSG for sponsoring the event and the BCS EWICS Service, and in particularly Jutta Mackwell, for the online publication services and finally we like to thank Guido Zuccon for compiling the proceedings.

ORGANIZATION

General Chairs

Massimo Melucci, University of Padua
Ricardo Baeza Yates, Yahoo! Research

PC Chair

Leif Azzopardi, University of Glasgow

Program Committee

Alex Bailey, Google
Andy MacFarlane, City University
Bettina Berendt, KU Leuven
David Elsweiler, University of Erlangen-Nuremburg
David Losada, University of Santiago de Compostela
Gareth Jones, Dublin City University
Ian Ruthven, University of Strathclyde
Jaap Kamps, University of Amsterdam
Juan Manuel Fernandez, University of Granada
Jun Wang, University College London
Leif azzopardi, University of Glasgow
Maarten de Rijke, University of Amsterdam
Maristella Agosti, University of Padua
Mark Baillie, University of Strathclyde
Massimo Melucci, University of Padua
Micheal Oakes, University of Sunderland
Milad Shokouki, Microsoft Research
Monica Landoni, University of Strathclyde
Mounia Lalmas, University of Glasgow
Norbert Fuhr, University of Duisburg Essen
Paul Thomas, CSIRO
Udo Kruschwitz, Univesity of Essex
Susanne Oehler, University of Glasgow
Vinay Vishwa, Microsoft

Table of Contents

Keynote

Papers

Posters

Readersourcing: Scholarly publishing, peer review, and barefoot cobbler's children

Stefano Mizzaro
Dept. of Mathematics and Computer Science
University of Udine
mizzaro@dimi.uniud.it

ABSTRACT

In this talk, I will start from an introduction to the field of scholarly publishing, the main knowledge dissemination mechanism adopted by science, and I will pay particular attention to one of its most important aspects, peer review. I will present scholarly publishing and peer review aims and motivations, and discuss some of their limits: Nobel Prize winners experiencing rejected papers, fraudulent behavior, sometimes long publishing time, etc. I will then briefly mention Science 2.0, namely the use of Web 2.0 tools to do science in a hopefully more effective way. I will then move to the main aspect of the talk. My thesis is composed of three parts.

1. Peer review is a scarce resource, i.e., there are not enough good referees today. I will try to support this statement by something more solid than the usual anecdotal experience of being reject because of bad review(er)s — that I'm sure almost any researcher has experienced.
2. An alternative mechanism to peer review is available right out there, it is already widely used in the Web 2.0, it is quite a hot topic, and it probably is much studied and discussed by researchers: crowdsourcing. According to Web 2.0 enthusiasts, crowdsourcing allows to outsource to a large crowd tasks that are usually performed by a small group of experts. I think that peer review might be replaced — or complemented — by what we can name Readersourcing: a large crowd of readers that judge the papers that they read. Since most scholarly papers have many more readers than reviewers, this would allow to harness a large evaluation workforce. Today, readers's opinions usually are discussed very informally, have an impact on bibliographic citations and bibliometric indexes, or stay inside their own mind. In my opinion, it is quite curious that such an important resource, which is free, already available, used and studied by the research community in the Web 2.0 field, is not used at all in nowadays scholarly publishing, where the very same researchers publish their results.
3. Of course, to get a wisdom of the crowd, some readers have to be more equal than others: expert readers should be more influential than naive readers. There are probably several possible choices to this aim; I suggest to use a mechanism that I proposed some years ago, and that allows to evaluate papers, authors, and readers in an objective way. I will close the talk by showing some preliminary experimental results that support this readersourcing proposal.

Disclaimer: This talk might harm your career; don't blame me for that.

Keywords: Scholarly Publishing, Science 2.0

Evaluation of User Comprehension of a Novel Visual Search Interface

Adrian O'Riordan
Computer Science, Sir Robert Kane Building, University College Cork, Cork, Ireland
a.oriordan@cs.ucc.ie

Advances in Web client technology now enable a richer interface thus supporting the application of visualization to the Web search process. This paper describes a user-centred evaluation of a novel visual search interface. The interface is designed to transparently support both Boolean and frequency models of information retrieval. The interface displays queries and returned results in a single simple graphical representation called Search Sphere. In contrast to both approaches that represent only queries visually and approaches that represent only results visually we integrate queries and result sets in a single visualization. We performed a user-centred study to evaluate both search query and result set comprehension. Comprehension in this context is defined as a user's correct interpretation of a given query and also his/her ability to select relevant documents from the result set visual. The feedback of this pilot study will be used to guide the development of a second prototype consisting of a rich client visual Web search interface.

Keywords: Visual Search Interfaces, Extended Boolean Search, Query Comprehension, User-centred Evaluation

1. INTRODUCTION

Visual search interfaces and the visualization of information spaces are active areas of research but are not yet to be widely used for Web search. The work described here is at the intersection of the fields of usability engineering, information retrieval and information visualization. We describe a visual search interface that transparently supports various models of information seeking and retrieval. The underlying retrieval model is left implicit so that the interface will be suitable for a broad spectrum of systems and users [9]. We evaluated a specific system aspect, user comprehension, using a low-fidelity prototype [25] to gain a micro-view of the requirements for interactive visual search. A user-centred evaluation was carried out that will feed into the design of a second fully functional prototype. User-centred evaluation of information seeking systems is advocated by Marchionini as a means of tackling the user-centred paradox, the fact that we cannot know "how users can best work with systems until the systems are already built" [22]. We sought to increase understanding of how visual information presentation of the search query and results impacts on user comprehension of same. User comprehension is defined as a user's correct interpretation of a visual query representation with regard to a user's ability to select the most relevant documents from the corresponding result set visualization.

The paper is organized as follows. Section 2 contains relevant background material. Section 3 introduces the Search Sphere visualization and Section 3.1 discusses specific issues related to user comprehension of same. Section 4 contains the evaluation and suggestions for progression. Section 5 describes related work. The paper finishes with a conclusions section.

2. BACKGROUND ON VISUAL SEARCH

Numerous techniques have been developed for visualizing the information space of a document collection. For example using "points of interest" visualizations where forces of attraction are used to position a document's location in a display [12]. Visualization of document clusters [21][38] organise results with spatial layouts. Another approach to visual search is to explicitly show query-document associations, as in Hearst's TileBars visualization [11]. Much of this work focuses on visualizing the document collection or query results and not the querying process itself. Our work focuses on graphical querying and the related results presentation.

Graphical approaches to query formulation are an alternative to the dominant text based method. Intended benefits are tied to the properties of direct manipulation [30]. Graphical approaches attempt to reframe querying and the search process as direct manipulation. Early work on graphical querying focused on visual representations of Boolean expressions as applied to querying relational databases [23]. These ideas have also been applied to text search where Boolean queries are represented visually and again evaluated as set operations. Efforts were motivated by the difficulty users have in formulating Boolean requests textually [9][31]. Young and Shneiderman used the visual form of boxes, representing terms, arranged in sequence and parallel to represent term disjunctions and conjunction respectively [7]. A prevalent visual scheme for Boolean search has search terms represented as circles or ovals, akin to the sets of Venn diagrams in mathematics. In this representation, the intersection of circles corresponds to the conjunction of the corresponding terms. Note that the standard Boolean retrieval model has no provision for ranking results. This and other limitations of existent systems will be further highlighted in Section 3. VQuery is a system that implemented these ideas to provide an interface to digital libraries [17]. Shneiderman conceptualized the notion of dynamic queries, a graphical query mechanism aimed at both novices and experts [1]. Dynamic approaches to querying allow users to create, move, and remove these circles in order to formulate and reformulate queries, ideally with rapid feedback of the result sets [30].

3. SEARCH SPHERE VISUALIZATION

Users searching for or trying to find information are faced with a number of challenges during the search process such as precisely stating their needs and effectively using the search interface. Two of the principal tasks or activities involved are query formulation and result selection. Various and diverse models of interactive information search and seeking have been proposed [26][20] that go beyond classical information retrieval models [23, 5] but all include the translation of an information need [35] into a form the machine can understand and necessitate the user interpreting the search results. We focus in particular on *informational needs* [7], where the intent is to acquire information assumed to be present. A query is a precise statement of an information need that a computer can process. Seach is usually an iterative process involving successive query reformulation. We took a user-centred approach in that we started design from the perspective of users' information needs and not retrieval models and carried out an early user evaluation to learn more about the requirements.

In practice, most current approaches, including the popular search engines, separate the query formulation from the display of results and results selection. In terms of screen real estate, a query (composed of search terms) is often constructed in one area, and the results displayed as a list in a separate area of the page. Search terms can be composed using query operators such as the Boolean operators. In these systems, even where results are displayed visually, as in the aforementioned KartOO visual search interface, the query is still entered in a separate search box. This separation divides user focus; eye tracking experiments of problem solving involving diagrams with separated textual annotations have found "frequent alternating fixations on specific pictorial and textual parts" [4]. This problem can be viewed as an instance of what Allen calls a user's failure of perception [3]. Thus cognitive factors influence search performance. The Search Sphere visualization was kept as simple as possible to avoid both cognitive load during user familiarization and information overload during use. This preference for a simple scheme is supported by research, for example a comprehensive meta-analysis found that "users tend to perform better with simple visual-spatial interfaces" [8].

In Search Sphere, search terms are represented as rings. Documents deemed to be relevant to a search term as shown within a ring. Currently what is displayed is an elided document name (first 25 characters). The Search Sphere contains one or more semi-transparent rings each corresponding to a search term. The terms themselves are placed at a position on the circumference of the Search Sphere, which we call the spoke which is also the centre point of its ring. Intersection of rings represents the conjunction of the corresponding terms in the query; non-intersection represents disjunction. In contrast to Venn diagram approaches such as Jones [17], results are also shown as part of the same visual representation. (In Jones' work results were displayed conventionally as a separate list.) Unlike in pure Venn diagram visualizations, these rings are not mathematical sets; the position of a document in a ring is significant and gives a measure of relevance. In Search Sphere the distance of a document's location from the corresponding spoke gives the relevance of that document to a search term. Gradient shading is used as a visual cue.

Search Sphere is not tied to any particular search system rather it is conceived as a front end to a search engine. The examples given here employed Google Search as the backend search. Multiple backend searches for the multiple queries, or points of interest [24], are required to produce the results. This is explained with an example below. Figure 1(a) shows an example with a single search term, New York. The results displayed are Google Search results where the distance of the document location from the spoke corresponds to the rank in results. The radial position is defined as the position of a document with respect to the line connecting the spoke to the centre of the Search Sphere. In this example the radial positions are not significant and where placed randomly. If a search engine supports result set clustering or document-document similarity measures this information can be used to cluster/position results in this representation.

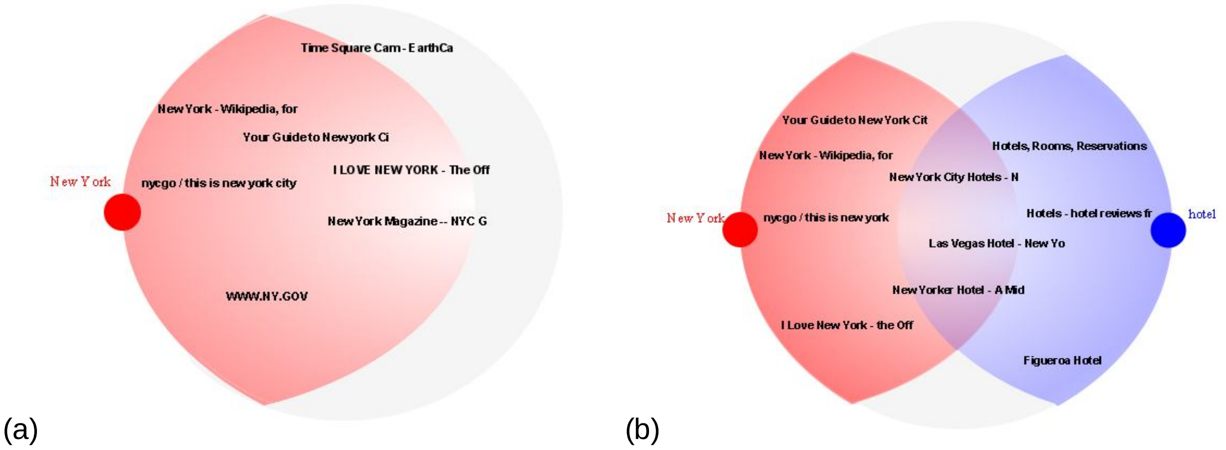

Time Square Cam - EarthCa
New York - Wikipedia, for
Your Guide to Newyork Ci
I LOVE NEW YORK - The Off
New York
nycgo / this is new york city
New York Magazine -- NYC G
WWW.NY.GOV
(a)
Your Guide to New York Cit
Hotels, Rooms, Reservations
New York - Wikipedia, for
New York City Hotels - N
New York
nycgo / this is new york
Hotels - hotel reviews fr
hotel
Las Vegas Hotel - New Yo
New Yorker Hotel - A Mid
I Love New York - the Off
Figueroa Hotel
(b)

FIGURE 1: (a) Single search term and (b) Two search terms

In Figure 1(b) the first example is developed further by adding a second search term, hotel. Now you can see that results for New York U hotel, New York \ hotel, and hotel \ New York, where U is the set intersection operator, and \ the set-theoretic difference or relative complement operator. In addition, the distances of the document titles from the spokes connote information. The documents in the central intersection area are the top results returned by Google from the query "New York hotel". The relevance of documents with respect to individual search term queries could be used to position the results slighted closer to either spoke; note that this information is not readily available from Google Search. Clustering information, if available, could be utilized to place results in radial positions, but the clusters would be tempered by the need to represent distance from a spoke which should take precedence over cluster location. Spoerri argues that the lack of a strong visual cue for relevance makes it hard for users to decipher the visual maps of many document-set visualization schemes [34].

3.1 Interpretation of Visual Metaphor

This simple visual can express Boolean queries, term frequency approaches to search [29], hybrids of both, and also clustered results. By hybrid approaches we are referring to models of information search and retrieval that incorporate aspects of frequency models and set based (for example Boolean) search. One well-know hybrid is the p-norm Extended Boolean Model [28] which is an parameterized intermediary between the vector processing model and Boolean query processing and collapses to either depending on parameter values. Strict interpretations of Booelan queries have been found not to be compatible with the user's interpretation [28]; neither is there a provision for term weight assignment or non-binary relevance. Hence a preference for "soft" interpretations of Booean requests in system implementations. Search Sphere has no explicit parameter, catering for either interpretation depending on the underlying mechanisms.

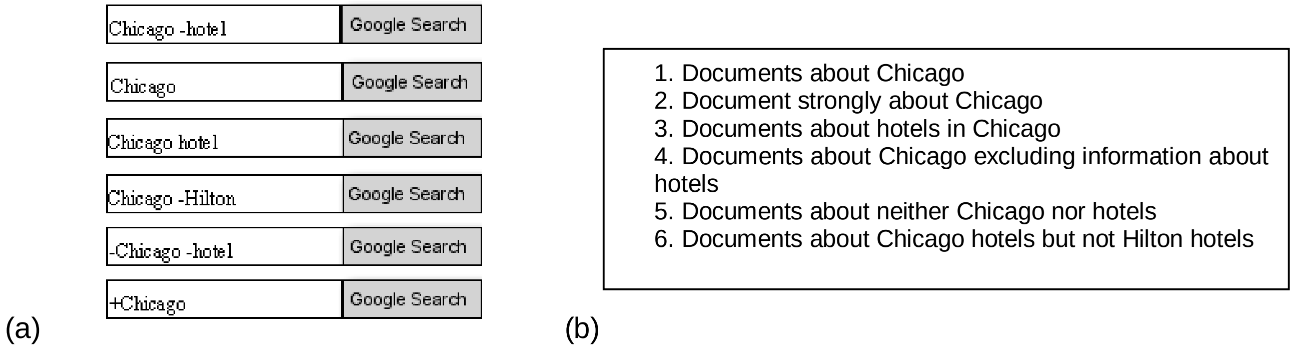
A problem with representing each term as a circle in Venn diagram based representations is that it is awkward to represent the conjunction of more than three terms. For example, the VQuery system circumvents this by allowing the assignment of multiple query terms to a single circle [17]. While Venn diagrams can represent four or more overlapping sets, the diagrams get cluttered. We do not believe this is a major weakness since user queries tend to be very short and rarely will involve intersections of four or more terms. Search log analysis has consistently highlighted low mean query length and a low mean percentage of queries containing an intersection expression; one study gave 2.2 and 8 respectively [18]. Additionally, inventive techniques for the graphical display of more complex expression and improving visual clarity have been developed [16].

The predominant list-based representation gives few cues to support the important task of deciding on a given result's relevance besides the list ordering. (Displaying document surrogates is one way of tackling this.) In contrast, each part of the Search Sphere can be viewed to represent a different aspect of a search. For example, the results listed in the area representing the overlap of two terms relate to the logical AND of those terms. Only a small number of the most relevant results for each such aspect are displayed so as not to clutter the view causing visual crowding. We restrict displayed results to seven for each query term. We thus subdivided the problem space of visual search into two connected parts: One part of the problem is the selection of search terms and the positioning of rings to form more complex query expressions. The other part is the user interpretation of a diagram vis-à-vis comprehending the query and the corresponding map of results. We evaluated the later.

4. EVALUATION OF USER COMPREHENSION

We conducted an early lifecycle small-scale user study to ascertain estimates of user comprehension of Search Sphere. College undergraduate students were given search tasks to perform with the aim of determining 1) error rates and 2) preference for this type of visual search as opposed to familiar text-based search. It was the expectation that most participants were familiar with Web search. The anonymous evaluation took the form of a questionnaire that was given to undergraduate college students via prearranged contact ensuring a very high response rate. There was a time limit of 12 minutes. All of the questions were factual except for one question on preferences. We obtained results from 54 persons, from two disciplines: 31 from Computer Science (CS) students and 23 from Occupational Therapy (OT) students. These two groups were chosen because both would have previous exposure to Web information retrieval systems but would have used different search tools in the process.

The questions were grouped into four sections: 1) participant's search experience, 2) Google Search tasks, 3) visual search tasks, and 4) comparison of Google Search and the visual search approach. The goals of the study were presented up-front at the start. The first section contained general questions about respondents' personal profile (age, gender) and familiarity with various search engines and search methods/techniques. Section two focussed on students' ability to comprehend Google search queries of varying complexity. Boolean querying was briefly explained with examples. The Google Search interface was presented as one would see it on screen, with queries filled in for six different searches; see Figure 2(a). The six information needs for these queries were also listed but in a different order; see Figure 2(b). Participants were asked to match the queries and information needs.


Chicago -hotel
Chicago
Chicago hotel
Chicago -Hilton
-Chicago -hotel
+Chicago

1. Documents about Chicago
2. Document strongly about Chicago
3. Documents about hotels in Chicago
4. Documents about Chicago excluding information about hotels
5. Documents about neither Chicago nor hotels
6. Documents about Chicago hotels but not Hilton hotels

FIGURE 2: Text search evaluation: (a) Google Search expressions and (b) Information needs

The third section introduced Search Sphere and asked the respondents to identify relevant documents for the same information needs as in section two. Included was a brief discussion of the visual metaphor. Instead of formulating queries, respondents had to choose documents that they deemed to be most relevant. The graphic in Figure 3 accompanied the questions. Note that documents were represented as simply doc1, doc2, etc. so that the titles didn't bias selection. The final section asked respondents which scheme they preferred, if any, and why.

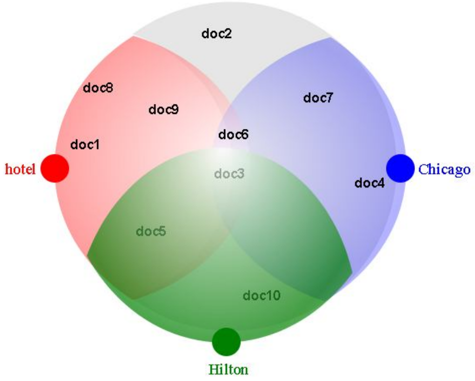


FIGURE 3: Visual search example in question

4.1 Results and Analysis

All the OT students had used Google Search, many Yahoo! Search and some specialized medical/health search systems (classified as non-Web Search here) but few had used advanced query techniques; see Table 1. The numbers in Table 1 are the number of students from the total of 23 who have experience of this search engine/ type of search. The average number of correct results for the visual search was slightly higher than for Google Search; see Table 2. The numbers in columns are the numbers who got the questions correct. The normalized average is the average of column seven scaled to a maximum for 1. The scaling allows comparison with the CS Group results below. Blank answers were interpreted as being incorrect. The low value for visual search for question six was partly down to respondents running out of time.

TABLE 1: Search history (OT Group: 23 Total)

Google Search	23	*Yahoo! Search*	17	*Live Search*	10
Any Other	5	*Advanced Google Page*	11	*Non-Web Search*	16
Boolean AND/OR	2	*Term Exclusion*	1	*Phrasal Search*	8
Other Techniques	4				

TABLE 2: Correct interpretations (OT Group: 23 Total)

	Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Average*	*Normalized Average*
Google	22	16	20	13	8	15	15.7	0.68
Visual	18	20	14	16	18	9	15.8	0.69

As expected the CS students had on average used more search engines and search techniques. They performed better than the OT group on the Google search tasks (0.75 as opposed to 0.68 for the normalized average) and they performed even better on the visual search tasks (0.86 as opposed to 0.69 for the normalized average). In particular both the OT and CS groups performed much better at the visual search for Q5 which contained two negations.

TABLE 3: Search history (CS Group: 31 Total)

Google Search	31	*Yahoo! Search*	24	*Live Search*	9
Any Other	11	*Advanced Google Page*	21	*Non-Web Search*	18
Boolean AND/OR	13	*Term Exclusion*	9	*Phrase Search*	21
Other Techniques	10				

TABLE 4: Correct interpretations (CS Group: 31 Total)

	Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Average*	*Normalized Average*
Google	27	18	30	25	18	22	23.3	0.75
Visual	31	29	22	29	28	21	26.7	0.86

Overall accuracy is higher for the visual search approach with fewer mistakes in particular when negation is involved. Both groups showed the same inclination with the difference wider for the more search-savvy CS group. These preliminary results are to feed into the development of an interactive Web search interface.

4.2 Discussion and Second Prototype Development

These are preliminary findings based on a mock-up prototype. In carrying out the system design and evaluation as few assumptions as possible where made about user's search behaviour. For example we did not restrict the interface or the evaluation to cater for a particular type of informational need. User studies have revealed users' search behaviour to be both diverse and unpredictable [32]. Constraining a future study to focus the examination on particular seeking tasks may elucidate more about visual search. Also the quality of a visualization is influenced by multiple factors such as expressiveness, appropriateness and effectiveness and more tests are required [6]. A second generation prototype is now being developed in Silverlight, a rich client application development environment created by Microsoft for the Web. This will enable experiments to be performed on query construction as well as comprehension. An interactive interface will allow the creation, movement and re-sizing of rings with immediate feedback of the changes in results. Cognitive research on the graphical query construction process found that two different search patterns are used to express graphical queries, neither of which is pervasive [36]. Set assembly is the classical intersecting sets in classical Venn diagrams. Set refinement is the iterative creation of successive subsets. They concluded that directly manipulable graphical interfaces that give users control over the creation and placement of sets consequently warranted investigation.

We are also interested in investigating cognitive factors such as how individual differences in visual-spatial ability may lead to performance differences in information search. Research has already shown there to be significant cognitive differences in text-based search [2]. It is highly likely that individuals will have a preferential leaning towards a particular approach. One issue in the preliminary results above is that users took longer to process the visual queries. This is an issue that will be addressed in the creation of a second prototype.

5. RELATED WORK

Section 5.1 reviews related usability studies of relevance and Section 5.1 summarizes emerging Web-based search visualizations.

5.1 Visual Search Usability Studies

Hertzum and Frokjaer [13] compared Boolean retrieval using a text-based and a Venn diagram representation and found that the later was superior in terms of mean time required to formulate a query and error rate. Grokker has a complicated visual interface which may help explain why one usability study found low user satisfaction [27]. User studies of set assembly (term disjunction) and set refinement (term conjunction, negation) in Venn diagram representations have revealed that users make fewer errors constructing queries than with text based interfaces [15]. The formulation of Boolean queries textually is surprisingly error-prone [31]. With regard to the specific issue of search results presentation, Hoeber and Yang presented a comparison of conventional list-based representations with interactive visual representations [14]. The results of this small user study indicate that users find the interactive visual approach more effective and satisfying.

5.2 Related Visualizations

Web technology has now developed to the point where these dynamic features can be added to a Web search interface. Rich client applications using technologies such as AJAX, Flash and Silverlight allow interaction in the browser. Many rich client search interfaces have been built in the last few years but only have a tiny share of the search market. The Flash-based KartOO (http://www.kartoo.com) is one example of a visual Web search interface but like many such systems focuses on the visualization of search results; the initial search terms are entered via a conventional search box. An example of a Web search interface that does visualize the query is Boolify (http://boolify.org), a Web-based educational tool enabling learners to construct Boolean queries out of puzzle pieces with a simple drag-and-drop interface. Search terms as well as AND, OR and NOT operators are represented as pieces that can be knitted together as in a jigsaw puzzle. Sortfix (http://sortfix.com/) is another search interface supporting drag-and-drop of search terms aimed at children. TwitterVenn (http://www.neoformix.com), built using the Processing programming language, allows the visualization of search terms in Twitter "tweets" as Venn diagrams. Ujiko (http://www.ujiko.com) and eyePlorer (http://www.eyeplorer.com) both use the radials of a Flash-based circle representation to guide query expansion.

The CircleSegmentViews system shares many characteristics with Search Sphere such as visual representation of AND and OR operators, use of radial information, colour coding but maps the results sets using meta-data rather than query terms. Klein and Reiterer carried out a user-centred study of visual query formation using this system [19]. Location, colour and distance are utilized in the WebStar model; a "display sphere" contains subjects which define interest centres [39]. Mooter (http://www.mooter.com) groups results in terms of themes so users can pursue the theme(s) that they are interested in. In the InfoCrystal system, in addition to supporting visual representations of Boolean queries, users can "assign relevance weights to the concepts and formulate weighted queries by interacting with a threshold slider" [33]. A system that also shares a number of features with Search Sphere is MetaCrystal [34], although MetaCrystal does not directly represent Boolean queries. With MetaCrystal a direct manipulation interface enables users to iteratively compose and edit meta-searches. MetaCrystal uses a "bull's eye" layout to enable users to visually compare the search results of multiple retrieval engines. Shape (size), colour, orientation and proximity are used to visually organize different search engine results in a circle. Whereas Search Sphere represents document relevance in terms of the distance from a spoke, MetaCrystal has a rank spiral, where most relevant results are closer to the centre. Sparkler is a visual meta-search engine using star plots where documents map to a position on a spoke based on relevance [10]. Grokker (www.groxis.com) is a visual meta-search engine that uses nested colour-coded circles or rectangles to visualize groupings of search results. Results are organized into multi-level categories. Subcategories are represented as smaller circles within their containing category. Filters can be applied, via sliders to restrict the search by source, domain, date, and information density.

6. CONCLUSIONS

The trend towards interactive Web interfaces will lead to increasing use of visual metaphors in search and information exploration. While research in visual user interfaces has already provided valuable lessons, more experimentation is needed to meet the requirements of visual interactive Web search. We carried out an evaluation of a novel search visualization building on work carried out previously. We assessed user query comprehension and results selection and compared against the dominant text based Web search approach as exemplified by Google Search. Results encouragingly showed that comprehension error rates were lower and preference stronger for the visual approach but the information load of processing a complex interface is a vital constraint. This will feed into the creation of a second interactive prototype which is underway.

ACKNOWLEDGMENTS

D. O'Byrne commented on a draft of the evaluation questionnaire. C. Hoare helped run the usability tests. Thanks also to F. Lane and the anonymous reviews for helpful comments.

REFERENCES

[1] Ahlberg C., Williamson, C. and Shneiderman, B. (1992) Dynamic queries for information exploration, *Proceedings of the SIGCHI*, Monterey, CA, 619-626.
[2] Allen, B.L. (1992) Cognitive differences in end-user searching of a CD ROM index. *Proceedings of ACM SIGIR.* Copenhagen, Denmark, 298-309.
[3] Allen, B.L. (1996) *Information tasks - toward a user-centered approach to information systems*. San Diego: Academic Press.
[4] Bertel, S. (2006) Visual focus in computer-assisted diagrammatic reasoning. In *Diagrammatic Representation and Inference*: LNCS 4045, 241-243. Springer Berlin, Heidelberg.
[5] Baeza-Yates, R and Ribeiro-Neto, B. (1999) *Modern Information Retrieval.* Addison Wesley, Harlow, England.
[6] Bederson, B.B. and Ben Shneiderman, B. (2003). *The Craft of Information Visualization: Readings and Reflections*. Morgan Kaufmann, San Fransisco, CA.
[7] Broder, A (2002) A taxonomy of web search. *ACM SIGIR Forum*, **36**(2), 3-10.
[8] Chen, C. and Yu, Y. (2000) Empirical studies of information visualization: A meta-analysis. *Int. J. Human-Computer Studies,* **53**, 851-866.
[9] Chowdhury, G.G. (2004) *Modern Information Retrieval, 2nd ed.*, Facet Publishing.
[10] Havre, S., Hetzler, E., Perrine, K., Jurrus, E. and Miller, N. (2001) Interactive visualization of multiple query results. *Proceeding of IEEE Information Visualization Symposium*, San Diego, CA.
[11] Hearst, M. (1995) TileBars: Visualization of term distribution information in full text information access, *Proceedings of SIGCHI*, Denver, CO, 59-66.
[12] Hemmje, M., Kunkel, C. and Willett, A. (1994) LyberWorld: A visualization user interface supporting fulltext retrieval. *Proceedings of ACM SIGIR*, Dublin, Ireland, 249-259.

[13] Hertzum, M. and Frøkjær, E. (1996) Browsing and querying in online documentation: A study of user interfaces and the interaction process. *ACM Transactions Computer-Human Interaction*, **3**(2), 136-161.

[14] Hoeber, O. and Yang, X.D. (2006) A Comparative user study of web search interfaces: HotMap, Concept Highlighter, and Google. *Proceedings of Web intelligence*, Washington, DC, 866-874.

[15] Jansen, B.J., Spink, A. and Saracevic, T. (2000) Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management,* **36**, 207-227.

[16] John, C., Fish, A., Howse, J., Taylor, J. (2006) Exploring the notion of clutter in Euler diagrams, *Diagrammatic Representation and Inference:* LNCS **4045**, 267-282.

[17] Jones, S. (1998) Dynamic query result previews for a digital library. *Proceedings of ACM Conf. Digital Libraries*, Pittsburgh, PA, 291-292.

[18] Jones, S., McInnes, S. and Staveley, M.S. (1999) A Graphical user interface for Boolean query specification. *International Journal on Digital Libraries*, **2**(2-3), 207-223.

[19] Klein, P. and Reiterer, H. (2005) The CircleSegmentView - A visualization for query preview and visual filtering, *Proc. SPIE*, **5669**.

[20] Kuhlthau, C.C. (2006). Kuhlthau's information search process. In K. Fisher, S. Erdelez, & L. McKechnie (Eds.), *Theories of Information Behavior*, 230–234, Information Today, New Jersey.

[21] Lagus, K., Kaski, S. and Kohonen, T. (2004) Mining massive document collections by the WEBSOM method. *Information Sciences*, **163**(1-3), 135-156.

[22] Marchionini, G. (1995) *Information Seeking In Electronic Environments*. Cambridge University Press, Cambridge, UK.

[23] Michard, A. (1982) Graphical presentation of Boolean expressions in a database query language: design notes and an ergonomic evaluation. *Behavior and Information Technology* **1**(3) , 279-288.

[24] Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B. and Williams, J.G. (1993) Visualization of a document collection: The VIBE system. *Information Processing and Management,* **29**(1), 69-81.

[25] Petrelli D. (2008) On the role of user-centred evaluation in the advancement of interactive information retrieval. *Information Processing and Management*, **44**(1), 22-38.

[26] Pirolli, P. (2007) *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press, Oxford, UK.

[27] Rivadeneira, W. and Bederson, B. (2003) A study of search result clustering interfaces: Comparing textual and zoomable user interfaces, ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts-Bibliography/2003-36html/2003-36.htm.

[28] Salton, G., Fox, E.A. and Wu, H. (1983) Extended Boolean information retrieval. *Communications of the ACM,* **26**(11), 1022–1036.

[29] Salton, G. and McGill, M.J. (1986) *Introduction to Modern Information Retrieval*. McGraw-Hill, London, England.

[30] Shneiderman, B. (1994) Dynamic queries for visual information seeking. *IEEE Software*, **11**(6), 70–77.

[31] Shneiderman, B. (1997) *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley, Harlow, England.

[32] Spink, A., Wilson, T., Ellis, D., Ford, N. (1998) Modeling users' successive searches in digital environments: A National Science Foundation/British Library funded study. *DLib Magazine*, April 1998.

[33] Spoerri , A. (1995) InfoCrystal. *Proceedings of ACM SIGIR*, Seattle, WA, p. 367.

[34] Spoerri, A. (2006) Visualizing meta search results: Evaluating the MetaCrystal toolset. *Proceedings of American Society for Information Science and Technology*. Austin, TX.

[35] Taylor, R.S. (1962) Process of asking questions. *American Documentation*, **13**, Oct., 391-396.

[36] Willie, S. and Bruza, P. (1995) Users' models of the information space: the case for two search models. *Proceedings of ACM SIGIR,* Seattle, WA, 205-210.

[37] Young, D. and Shneidermann, B. (1993) A Graphical filter/flow model for Boolean queries: An implementation and experiment. *Journal of the American Society for Information Science*, **44**(6), 327-339.

[38] Zamir, O. and Etzioni, O. (1999) Grouper: A dynamic clustering interface to web search results. *Proceedings of WWW*, Toronto, Canada, 1361-1374.

[39] Zhang, J. and Nguyen, T.N. (2005) WebStar: a visualization model for hyperlink structures. *Information Processing and Management*, **41**(4), 1003-1018.

Efficient Content-Based Information Retrieval: A New Similarity Measure for Multimedia Data

Christian Beecks   Thomas Seidl
Data Management and Data Exploration Group
RWTH Aachen University, Germany
{*beecks,seidl*}*@cs.rwth-aachen.de*

Abstract

Content-based information retrieval of multimedia data is a great and attractive challenge which raises numerous research activities. As multimedia data become ubiquitous in our daily lives, information retrieval systems have to adapt their retrieval performance to different situations in order to efficiently satisfy the users' information needs anytime and anywhere. To enhance the content-based multimedia information retrieval process, we propose an efficient similarity measure based on flexible feature representations. We define the similarity model for content-based information retrieval and show its feasibility on real world multimedia data. Furthermore, we briefly discuss future research directions which we plan to investigate.

1. INTRODUCTION

Information retrieval systems are ubiquitous in our daily lives. They support us to store, manage and access voluminous information based on the systems' underlying data. Thus, the primary aim of such information retrieval systems is to supply user's with relevant and useful information in order to satisfy their information needs. To this end, the system should perform the information retrieval process of a user-specific query in an effective and also efficient way.

Prominent examples of information retrieval systems are search engines in the World Wide Web. They enable us to do a search for interesting web pages, popular video clips, famous research papers, gorgeous images, and so on. Almost all multimedia information inside and outside the World Wide Web are made accessible via appropriate retrieval systems.

In order to fulfill retrieval and browsing tasks in the user's sense, information retrieval systems use different kinds of models to represent their accessible data through additional semantic and syntactic information. These information reflect the contents of the stored data and represent the core of each information retrieval process.

To perform the content-based information retrieval process in an efficient way, we propose a similarity model that extends the classic vector space model. We allow data objects to store multiple feature vectors and compare these weighted sets with our new similarity measure. Thus, the feature representation supports us to dynamically extract contents of multimedia data, whereas the new similarity measure efficiently compares these feature representations with each other.

We organize the paper's structure as follows: In Section 2, we briefly review existing works and backgrounds. We introduce our new similarity measure in Section 3 and give an impression of the first practical retrieval results in Section 4. We briefly describe future directions in Section 5 and conclude our paper in Section 6.

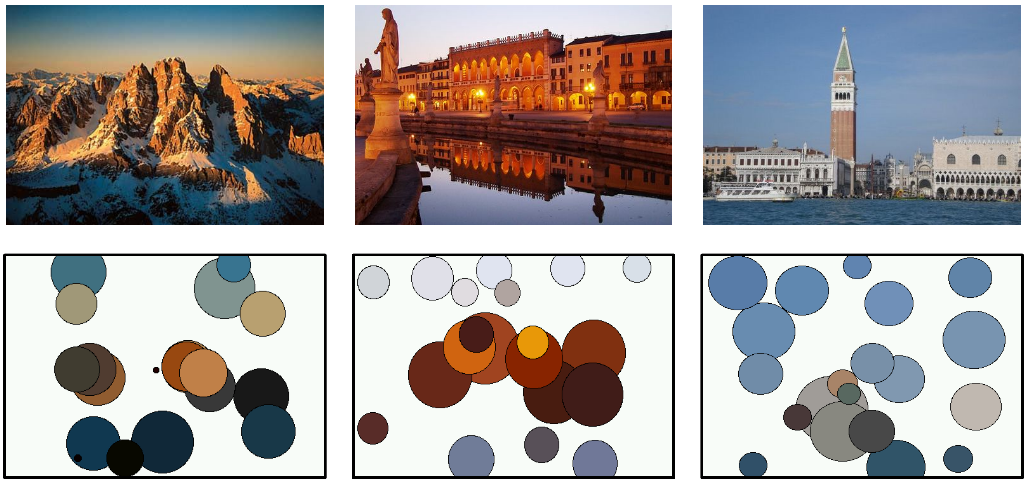


FIGURE 1: Three sample images in the top row with their signatures in the bottom row.

2. RELATED WORK AND BACKGROUND

The methodology of information retrieval covers a broad range of distinct interdisciplinary research areas like modeling and indexing data contents, searching information, evaluating retrieval results, designing user interfaces, and so forth. To get an impression and overview of this widespread field, we recommend the classic books of van Rijsbergen [1] and Salton et al. [2] and, for instance, the modern books of Beaza-Yates and Ribeiro-Neto [3] and Manning et al. [4].

Research in these different fields yields to a multitude of modern information retrieval systems. Each of the systems' core is the underlying information retrieval model. These models specify the way to store and access semantic and syntactic information of the data. Common classic models for information retrieval are the Boolean model, the vector space model [5], and the probabilistic model [6, 7]. These models are different in the way how they store the data and how they perform different user tasks and queries.

In the present work, however, we focus on an extended vector space model that is suitable for content-based multimedia information retrieval. Therefore, we store the contents of each data object in a weighted set of vectors and call this set a signature (cf. the work of Rubner et al. [8]).

Definition 1. (Signature)
Given a data object o, the corresponding signature S_o with length n^s is defined as

$$S_o = \{(v_i^s, w_i^s)\ |i = 1, \ldots, n^s\},$$

where n^s is the number of vectors $v_i^s \in \mathcal{R}^n$ with the corresponding weights $w_i^s \in \mathcal{R}^+$. The benefit of this feature representation, in contrast to the traditional vector approach, is the flexibility and adaptability of signatures to cover dynamic properties of vivid multimedia data. By extracting signatures from multimedia data objects, information retrieval systems are able to capture global as well as local object's properties to ensure effective content-based information retrieval processes.

Figure 1 shows an example of signatures from three different images. In this example, we extract signatures comprising 20 vectors with the underlying feature dimension equal to seven. These seven dimensions include two position components, three color components, one contrast component, and one coarseness component. In the figure, we visualize the vectors according to their position information with colored circles. The diameters of the circles reflect the weights of the vectors. As we see in this example, signatures represent a flexible and also compressed way to store data contents automatically without any additional semantic structure information.

To judge contents of multimedia objects based on their automatically extracted signatures, information retrieval systems compute similarities among signatures. Common similarity measures for signatures are the Earth Mover's Distance [8] or the Hausdorff Distance [9]. Both are applicable to multimedia data, because they use an adaptable ground distance to determine the distance between the signatures' vectors. Nevertheless, both distances limit the performance of modern information retrieval systems: Earth Mover's Distances have high computation times, whereas Hausdorff Distances suffer from the missing possibility to consider the weights of the signatures.

3. A NEW SIMILARITY MEASURE

In this section, we present our new similarity measure for multimedia information retrieval based on signatures. Therefore, we adapt the well-known concept of quadratic form distance measures [10, 11] from simple feature vectors to the more flexible feature representation of signatures. In order to compare two signatures, the challenging task consists in matching all of the vectors of both signatures among each other. To achieve this in the distance computation, we propose the following definition.

Definition 2. (Signature Distance)
Given two signatures $Q = \{(v_i^q, w_i^q)\ |i = 1, \ldots, m\}$ and $P = \{(v_i^p, w_i^p)\ |i = 1, \ldots, n\}$ of any length m, n, respectively, we define the Signature Distance measure SD as

$$SD(Q,P) = \sqrt{(\widetilde{w}_q - \widetilde{w}_p) \cdot A \cdot (\widetilde{w}_q - \widetilde{w}_p)^T},$$

where $\widetilde{w}_q \in \mathcal{R}^{n+m-k}$ and $\widetilde{w}_p \in \mathcal{R}^{n+m-k}$ are the extended permuted vectors with the same length which result from the signatures' weights $w_i^q \in \mathcal{R}^+$ and $w_i^p \in \mathcal{R}^+$ as follows:

$$\begin{aligned}
\widetilde{w}_q &= (\overbrace{w^q_{\pi(1)}, \ldots, w^q_{\pi(n-k)}}^{(n-k)\text{ weights of } Q}, \quad \underbrace{w^q_{\pi(n-k-1)}, \ldots, w^q_{\pi(n)}}_{k\text{ common weights}}, \quad 0, \ldots, 0) \\
\widetilde{w}_p &= \qquad\qquad (0, \ldots, 0, \quad \overbrace{w^p_{\pi(1)}, \ldots, w^p_{\pi(k)}}, \quad \underbrace{w^p_{\pi(k+1)}, \ldots, w^p_{\pi(m)}}_{(m-k)\text{ weights of } P}).
\end{aligned}$$

The extended permuted vectors $\widetilde{w}_q$ and $\widetilde{w}_p$ in the preceding definition consist of three blocks, according to the vectors v_i^q and v_i^p of the signatures. The first and the last block exclusively comprise weights from signatures Q and P, respectively, sharing no common vectors, whereas the block in the middle only consists of weights from Q and P sharing the same vectors. As a result, the permutation π aligns the weights of the signatures to each other and enables the multiplication with the similarity matrix $A \in \mathcal{R}^{(n+m-k)\times(n+m-k)}$ which is determined dynamically per distance computation. The dynamically generated similarity matrix A models the similarity among all vectors and depends on the compared signatures.

We give the following example to illustrate the distance computation of the proposed Signature Distance measure. For this purpose, we depict two signatures on the left-hand side in Figure 2. Signature Q consists of three vectors $v^q_{\pi(i)}$ which are depicted with the color light-blue, whereas signature P consists of two vectors $v^p_{\pi(i)}$ which are depicted with the color dark-blue. Both signatures share one common vector, namely $v^q_{\pi(3)}$ in signature Q and $v^p_{\pi(1)}$ in signature P. Thus, the extended permuted vectors have the total length of five and they share exactly one common component: the weights $w^q_{\pi(3)}$ and $w^p_{\pi(1)}$.

Based on the extended permuted vectors of the signatures which have to be compared, the similarity matrix is generated. In this example, the similarity matrix comprises three major blocks. The light-blue block and the dark-blue block determines the similarity among the vectors of signature Q and those of signature P, respectively. The overlapping block between the light-

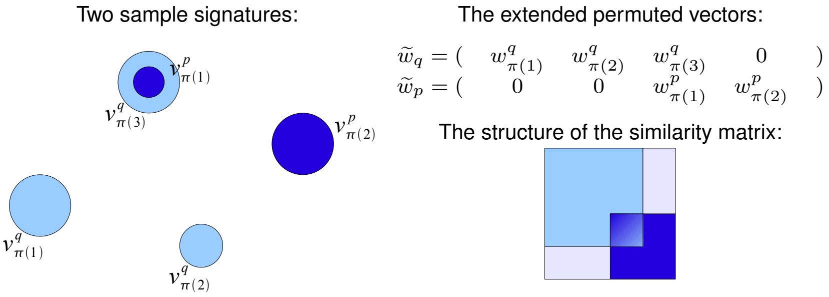

Two sample signatures:
$v^p_{\pi(1)}$
$v^q_{\pi(3)}$
$v^p_{\pi(2)}$
$v^q_{\pi(1)}$
$v^q_{\pi(2)}$
The extended permuted vectors:
$\widetilde{w}_q = (\; w^q_{\pi(1)} \; w^q_{\pi(2)} \; w^q_{\pi(3)} \; 0 \;)$
$\widetilde{w}_p = (\; 0 \; 0 \; w^p_{\pi(1)} \; w^p_{\pi(2)} \;)$
The structure of the similarity matrix:

FIGURE 2: Two sample signatures on the left-hand side, the extended permuted vectors and the structure of the similarity matrix on the right-hand side.

and dark-blue blocks models the similarity of the common vectors in both signatures. As in this example exists only one vector which appears in both signatures, this block consists of exactly one value. The other blocks of the similarity matrix model the similarities among the vectors of signature Q and signature P. As a result, the distance computation is finalized by multiplying the difference of the extended permuted vectors $\widetilde{w}_q$ and $\widetilde{w}_p$ with the generated similarity matrix according to Definition 2.

In order to generate the similarity matrix, we have to determine the similarities between two vectors. We recommend to choose a function which is inverse proportional to a distance function. Such a function guarantees the highest similarity value for the same vectors and lower similarity values for different vectors.

4. PRACTICAL RETRIEVAL RESULTS

In this section, we present the first practical retrieval results of our new similarity measure. For this purpose, we conducted several similarity queries on the *Wang* image collection [12] which consists of 1.000 images from 10 different themes. Based on a k-means clustering in the feature space, we extracted 20-dimensional signatures of the images including position, color, contrast, and coarseness information.

Figure 3 visualizes some of the retrieval results which are given below the three different query images with the most similar images on top of each column. The columns of each query show the ranked results of the proposed Signature Distance (*SD*), the Earth Mover's Distance (*EMD*), and the Hausdorff Distance (*HD*), respectively. The figure reveals that the results of the Signature Distance and Earth Mover's Distance continually have a higher quality compared to that of the Hausdorff Distance. Comparing the Earth Mover's Distance with the Signature Distance, we see that the latter has a slightly higher perceptual retrieval quality. Thus, we conclude that the computed similarities of the Signature Distance are well comparable to the ones of the Earth Mover's Distance. In order to verify the perceptual results, we list the aggregated mean average precision [4] values measured over 100 randomly chosen queries, based on the *Wang* image collection, in the following table.

	mean average precison
Quadratic Form Distance	0.51
Earth Mover's Distance	0.50
Hausdorff Distance	0.33

TABLE 1: Aggregated mean average precision.

In addition to the quality of the retrieval results, we also measured the computation times needed to generate the resulting rankings. As a result, the Hausdorff Distance has the lowest run-time of 23 milliseconds to generate the ranking. The Signature Distance has a run-time of 76 milliseconds, whereas the Earth Mover's Distance requires 261 milliseconds to finish the retrieval process. The

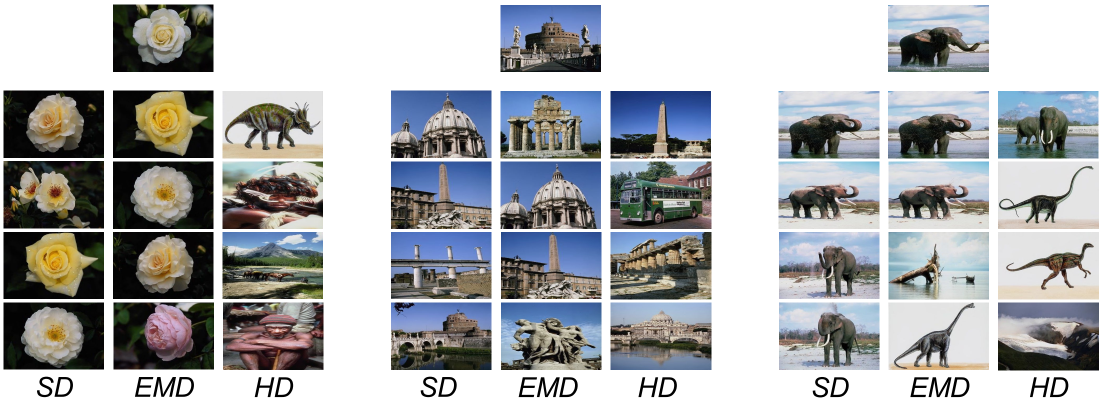


FIGURE 3: Three different query examples and their ranked results. For each query, the columns depict the results of the Signature Distance (*SD*), Earth Mover's Distance (*EMD*), and Hausdorff Distance (*HD*).

experiments were implemented in Java and the run-times were measured on Intel XEON E5345 CPU-based machine with 2.33GHz.

Combining the perceptual qualities and the run-times of the retrieval experiments, we summarize that our new Signature Distance advantages high retrieval qualities and comparatively low computation times. Both properties are fundamental for efficient content-based multimedia information retrieval.

5. FUTURE DIRECTIONS

For our future work, we plan to apply the proposed similarity measure on multimedia data to retrieve content-based information. Therefore, we identify the following future research directions on which we plan to contribute with our similarity measure:

In the field of region-based similarity search, we are interested in multimedia objects that are similar to a region of a given query object. Thus, instead of querying multimedia databases with complete signatures, we only use relevant components of query signatures to find those regions in the signatures containing relevant information of the objects. The arising question is, if this relevant components of the query signatures and of the signatures stored in multimedia databases can be determined automatically within the proposed similarity measure.

In addition to region-based similarity search, we plan to focus our research in the field of adaptable similarity search which considers the adaptation of the proposed similarity measure to different user preferences. In order to improve the retrieval quality of content-based similarity search, we plan to examine the properties of the underlying similarity matrix to capture those user preferences.

The third aspect that we want to consider is the content-based retrieval of heavily sized databases. As information retrieval is generally not restricted to a fixed size of the databases, we investigate on techniques to query voluminous data in an efficient way. To support the retrieval process, we plan to study approximation and indexing techniques of the proposed similarity measure.

6. CONCLUSION

We presented a new similarity measure based on flexible feature representations for efficient content-based information retrieval of multimedia data. We define and illustrate this similarity measure and show its feasibility and efficiency on real world multimedia data. As a result, we conclude that our new similarity measure combines a low run-time of distance computation with a high retrieval performance. Furthermore, we identify future research directions on which we plan to contribute with our new similarity measure.

REFERENCES

[1] C. Rijsbergen, *Information Retrieval*. Butterworth, 1979.

[2] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[3] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[4] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[5] G. Salton, *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1971.

[6] S. E. Robertson and S. K. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science*, vol. 27, no. 3, pp. 129–146, 1976.

[7] N. Fuhr, "Probabilistic models in information retrieval," *Computer Journal*, vol. 35, no. 3, pp. 243–255, 1992.

[8] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[9] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the Hausdorff Distance," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 15, no. 9, pp. 850–863, 1993.

[10] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and Effective Querying by Image Content," *Journal of Intelligent Information Systems*, vol. 3, no. 3/4, pp. 231–262, 1994.

[11] T. Seidl and H.-P. Kriegel, "Efficient User-Adaptable Similarity Search in Large Multimedia Databases," in *VLDB*, 1997, pp. 506–515.

[12] J. Wang, J. Li, and G. Wiederhold, "Simplicity: semantics-sensitive integrated matching for picture libraries," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 23, no. 9, pp. 947–963, 2001.

Web Personalization based on Usage Mining

Sharhida Zawani Saad
School of Computer Science and Electronic Engineering,
University of Essex,
Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK
szsaad@essex.ac.uk

Abstract

Personalized or recommender systems are a particular type of information filtering applications. User profiles, representing the information needs and preferences of users, can be inferred from log or clickthrough data, or the ratings that users provide on information items, through their interactions with a system. Such user profiles have been used, for example in iGoogle, to provide personalized recommendations to the users. A user model is a representation of this profile, which can be obtained implicitly through the application of web usage mining techniques.

Our work aims to develop Web usage mining tasks to model an intranet or local Web site recommender system. We will focus on the users activity on a university Web site, to customize the contents and structure the presentation of a Web site according to the preferences derived from the user's activity. The customization is based on an individual's user profile as well as a profile representing the collective interest of the entire user community, in this case all users accessing the Web site. The outcome will be personalized recommendations and presentation of a Web site with respect to the user's needs.

Keywords: Web usage mining, Recommender systems, Adaptive Web sites, Personalization

1. INTRODUCTION

The explosive growth of the Web has triggered an increasing demand of Web personalization systems. Personalized information technology services have become a ubiquitous phenomenon, taking advantage of the knowledge acquired from the analysis of the user's navigational behavior or usage data. Web usage mining (WUM) aims at discovering interesting patterns of use by analyzing web usage data. This method provides an approach to the collection and pre-processing of those data, and the construction of models representing the behavior and the interests of users. These models can be automatically incorporated into personalization components, without the intervention of any human expert (Girardi and Marinho, 2007).

Past research only focuses on how to meet consumers' information needs from the perspective of functional information, and they exclude most of the other information needs from their consideration (France et al., 2002). Research conducted by France et al. stated that the attempts of search engines and data mining technology to improve Web information search capabilities to match up various information needs has been limited. Our research aims to work on usage mining techniques to provide a user with personalized recommendations, by customizing the contents of a Web site with respect to the user's needs.

Let us use a university Web site as an example. Every user of that Web site (be it a student, a member of staff or an external visitor) will have different interests, the challenge is to tailor the presentation of the document collection according to each of these user's profiles. Obviously, this will only work if the user is in fact interested in having a personalised Web site. If not, then no user activity will be recorded and no personalisation takes place. The Web site appears unaltered.

While each user has individual interests, we can expect a lot of overlap. A newly registered student who is searching for the teaching timetable is not alone. A lot of other students share this information need and will have searched (and hopefully located) the appropriate documents on the Web site. Hence, a Web site presented according to the entire user community's information access activities is likely to make it easier for new users to find relevant documents quickly. Again, there is no need to enforce such customization, this can easily be switched off and the Web site appears as normal.

The point of the above example is that we are hoping to capture both user and community search/navigation trails to make the entire Web site adaptive and customize it according to the two profiles which capture these activities: user and community profiles.

Originally, the aim of Web usage mining has been to support the human decision making process (Baraglia and Silvestri, 2007). Thus, the outcome of the process is typically a set of data models that reveal knowledge about usage patterns of users. WUM typically extracts knowledge by analyzing historical data such as Web server access logs, browser caches, or proxy logs. Using WUM techniques, it is possible to model user behavior, and therefore, to forecast their future movements (Baraglia and Silvestri, 2007). The information mined can subsequently be used in order to personalize the contents of Web pages. Web personalization (WP) or recommender systems are typical applications of WUM. Although most of the work in Web usage mining is not concerned with personalization, its relationship to personalization issues has brought promising results (Girardi and Marinho, 2007). The knowledge discovered through the usage mining process serves as operational knowledge to personalization systems (Girardi and Marinho, 2007). Realizing the potential of WUM techniques to construct this knowledge, our proposed research aims to provide the personalization recommendations to the users, as an output from analyzing the user's navigational behavior or usage data.

2. RESEARCH QUESTIONS AND OBJECTIVES

We are conducting research designed to answer the following questions:

- How can Web usage mining techniques capture the behaviour and interests of users of a Web site?
- How can the captured knowledge be utilized to construct personalized recommendations on the content and presentation of the Web site?

The following objectives have been formulated to answer the research questions:

- To develop Web usage mining tasks based on some phases and applied techniques such as data collection, data pre-processing, pattern discovery and knowledge post-processing
- To construct models representing the behavior and the interests of individual users as well as a community of users from local Web sites or intranets
- To construct an integrated profile for any related patterns of profile based on the user and community profiles
- To construct personalized presentations guided by profiles (based on the integrated profiles).

3. RELATED WORK

Relevant work and field studies related to Web usage mining and personalization are discussed. Among the important subjects being discussed are adaptive Web sites, Web usage mining, recommender systems, and personalization, its previous work and how they relate to our research.

3.1. Adaptive Web Sites

Adaptive web sites automatically improve their organization and presentation by learning from user access patterns (Perkowitz and Etzioni, 1997). There is a lot of work going on about mining the users' clickthrough patterns and understanding query intent to improve search result sets, e.g. (Hu et al., 2009) and (Bayir et al., 2009). Adaptive web sites can make popular pages more accessible, highlight interesting links, connect related pages, and cluster similar documents together. Perkowitz and Etzioni discuss possible approaches to this task and how to evaluate the community's progress. The focus is either on customization: modifying web pages/site's presentation in real time to suit the needs of individual users; or optimization: altering the site itself to make navigation easier for all. One user's customization does not apply to other users; there is no sharing or aggregation of information across multiple user, and transformation has the potential to overcome both limitations (Perkowitz and Etzioni, 2000). In the Web domain, we observe a visitor's navigation and try to determine what page she is seeking, to offer the desired page immediately. Index page synthesis is a step towards the long-term goal of change in view: adaptive sites that automatically suggest reorganizations of their contents based on visitor access patterns (Perkowitz and Etzioni, 1999).

Our research will be focusing on customization: to modify Web pages/site's presentation to suit the needs of individual users. However, this will be based on the integrated profile, which combine both the user profile and community profile. Perkowitz and Etzioni introduce transformation as their approach, while our work will be using both customization (based on the user profile), and also transformation (which is related to the community profile). Compared to Perkowitz and Etzioni's work, this work will be fully automatic and does not involve any Web administrator / webmaster within the process. Apart from that, their previous work and motivation focus much on index page synthesis, and aims to automate the placement of new index pages at the Web site. Their work are not focusing on user profiles and they are looking for adaptive sites as the long-term goal of the research, while this work is focusing on adaptive system and personalization based on the user and community profiles. Compared to other previous works, the main difference is to adapt its content and presentation based on the integrated profile, for local Web site or intranet access. Apart from that, the system will not be relying on explicit user feedbacks or working with any webmaster control.

3.2. Web Usage Mining

During the last years, researchers have proposed a new unifying area for all methods that apply data mining to Web data, named Web mining (Kosala and Blockeel, 2000). Web mining is traditionally classified into three main categories: Web content mining, Web usage mining, and Web structure mining. Web usage mining aims at discovering interesting patterns of use by analyzing Web usage data (Girardi and Marinho, 2007). It is the process of applying data mining techniques to the discovery of usage patterns from Web data generated by user interactions with a Web server, including Web logs, clickstreams and database transactions at a Web site or at a group of related sites (Corsini and Marcelloni, 2006). (Kruschwitz et al., 2008) suggest to apply log analysis technique to electronic document collections and intranets, as search in this type of collections has attracted much less attention, but locating information within such collections can be as difficult as the open Web. Web usage mining is an excellent approach to make dynamic recommendations to a Web user, based on his/her profile in addition to usage behaviour. Usage patterns extracted from Web data have been applied to a wide range of applications (Srivastava et al., 2000).

Different modes of usage or mass user profiles can be discovered using Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user clickstreams stored in Web log files (Nasraoui et al., 2008). These profiles can later be harnessed towards personalizing the Web site to the user. The issue of Web mining is discussed in detail in (Eirinaki and Vazirgiannis, 2003). Apart from Web usage mining, user profiling techniques can be performed to form a complete customer profile (Eirinaki and Vazirgiannis, 2003). Our research aims to model personalization components based on Web usage mining techniques. Generally, the outcome would be the model of Web personalization systems that elaborate the Web usage information based on user's navigational behavior. The following section discusses Web recommender and personalization systems, and how they relate to our research.

3.3. Recommender Systems

Collaborative filtering and content-based filtering are two types of common recommender systems. Collaborative filtering (CF) is the process of filtering or evaluating items through the opinions of other people (Schafer et al., 2007). Ratings in a collaborative filtering system may be gathered through explicit or implicit means, or both. According to Schafer et al., collaborative filtering can predict what information users are likely to want to see, enabling providers to select subsets of information to display. In this way, collaborative filtering enables the Web to adapt to each individual user's needs. Content-based recommendation systems analyze item descriptions to identify items that are of particular interest to the user. A profile of the user's interests is used by most recommendation systems. (Pazzani and Billsus, 2007) describe two types of information for user profiles; a model of the user's preferences and a history of the user's interactions with the system. One important use of the history is to serve as training data for a machine learning algorithm that creates a user model. When the user explicitly rates items, there is little or no noise in the training data, but users tend to provide feedback on only a small percentage of the items they interact with.

Content-based filtering and collaborative filtering have long been viewed as complementary. Content-based filtering can predict relevance for items without ratings, while collaborative filtering needs ratings for an item in order to predict for it. Content-based filtering needs content to analyze, and collaborative filtering does not require content. While people find the quality of multimedia data (e.g., images, video, or audio) for web pages important, it is difficult to automatically extract this information (Schafer et al., 2007). More often collaborative filtering and content-based filtering are automatically combined, sometimes called a hybrid approach. Such systems generally use CF to try and capture features like quality that are hard to analyze automatically. Our research aims to provide personalized navigational information that elaborates the web usage information based on individual user and user community profiles. This could be done based on the user's activity or interaction through emails, searching or navigation behaviour. In this case, collaborative filtering may not be necessary apart from content-based filtering, to capture important usage patterns. The research direction is more into providing personalized recommendations by identifying the area of interests of a user, and to adapt with the community profile.

3.4. Personalization

Web usage mining has been suggested as a new generation of personalization tools which can address the shortcomings of more traditional systems such as manual decision rule systems, collaborative filtering systems and content-based filtering agents (Mobasher et al., 2000). The idea of Mobasher et al. is to use page view and session information to automatically obtain usage profiles. The WebPersonalizer System is presented, which uses two methods to discover usage profiles: computing session clusters and association rule discovery. Research done by (Albayrak et al., 2005) propose a multi-agent system composed of four classes of agents: many information extracting agents, agents that implement different filtering strategies, agents for providing different kinds of presentation and one personal agent for each user. The personal agent should constantly improve the knowledge about "his" user by learning from the given feedback, which is taken for collaborative filtering. (Teevan et al., 2005) suggest to re-rank the results provided by current search engines. They mention that implicit user feedback is more useful compared to explicit feedback. Our research is not going to personalize the results provided by search engines, but will focus on a personalized presentation of the Web site.

(Agichtein et al., 2006) introduce robust, probabilistic techniques for interpreting clickthrough evidence by aggregating across users and queries. They interpret post-search user behaviour to estimate user preferences in a real Web search setting. They show that by aggregating user clickthroughs across queries and users, they achieve higher accuracy on predicting user preferences. Agichtein et al. aim to improve Web search ranking. Our research is not looking at Web search scenarios, but at intranet search. Besides that, Agichtein et al. use queries from query logs (from major Web search engines) with search results that were manually rated, and user interaction data. Differently, our research will be based on the user's profile and community profile

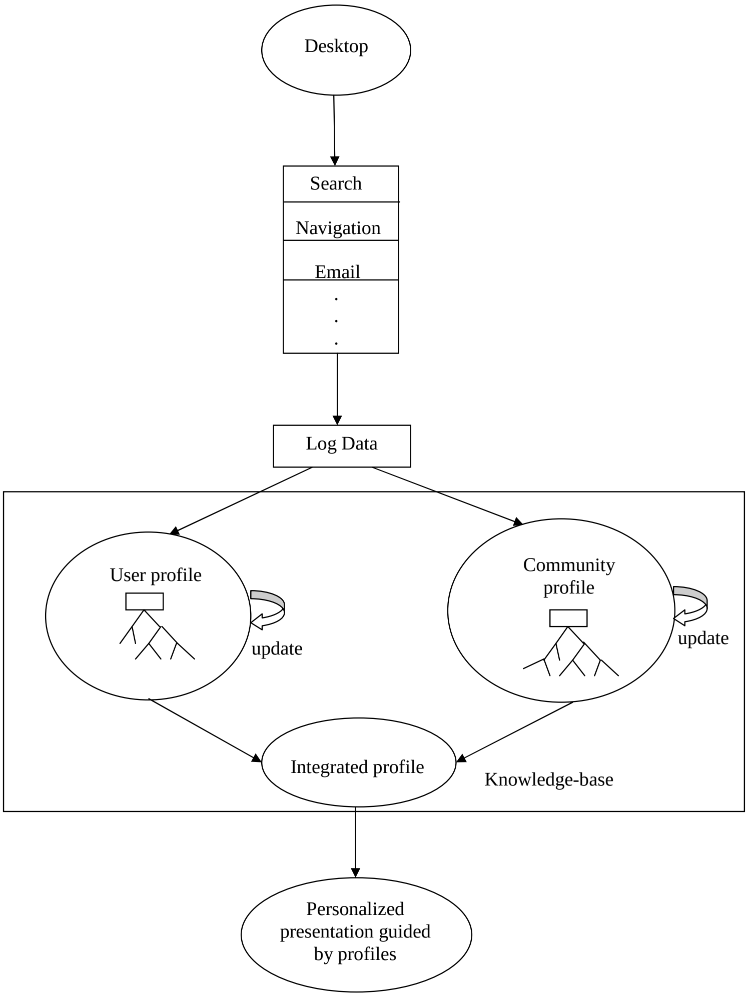


FIGURE 1: Preliminary architecture

to provide personalized recommendations, which will be done automatically and does not involve any manually rated results. A number of systems for personal search are available on today's PCs, including systems from Microsoft (desktop.msn.com), and Google (desktop.google.com). Research by (Cutrell et al., 2006) suggest Phlat, that combines keyword and property-value search, allowing users to find information based on whatever they may remember. Among problems and research directions being addressed is whether these designs can be extended to include 'non-personal' content (information sources of interest that users are not familiar with, such as news and intranet). Three research areas to be explored in personalization include recommendations, information filtering, and personalized presentation. These three areas are relevant but fairly big, therefore our future work will look into personalized presentation. This is to be done based on the user's profile from his searching/navigation activities, in combination with the community profile.

4. RESEARCH OUTLINE, DATA STRUCTURE AND EVALUATION

The system's preliminary architecture and some relevant data structures is shown in figure 1. The data flow is as follows:

- A user's activity is logged based on emails, searching and navigation behaviour.
- Web log analysis techniques are used to trace a user's behaviour or preferences.
- All the user's activities will be used to construct a user profile. The community profile is built in the same way without restricting it to a single user.
- An integrated profile combines user and community profiles in a uniform way.
- The outcome is a personalized presentation guided by profiles (based on the integrated profile).

Different types of data structure from Web page, email, text document and image can be treated in a uniform way. Each of them can be broken down into entities like title, descriptor or meta tags, content, unique identifier and contextual information. For example, contextual information of a Web page include Web documents under the same URL, while contextual information of an email include email of the same thread or information whether the email was read or written by the user.

We are at a very early stage of this research but we imagine that the outcomes will be useful in a number of different information access tasks including search, browsing and navigation. Ultimately, we are aiming at evaluating the developed techniques in some realistic Intranet settings and compare them against sensible baselines, e.g. standard search interfaces. In more detail, we plan to conduct task-based evaluations involving real users of a university Web site. The evaluations will have to address the questions whether information can be found in fewer interaction steps when using profiles, whether more relevant information can be found, whether the comparison of the two approaches (personalized versus baseline) shows differences in user satisfaction, etc. As a preliminary guide we use the evaluation paradigms developed for the TREC Interactive Track.

We do appreciate that one problem of such personalized search is the issue of ”consistency”, i.e. users expect consistency in a system's behaviour. We will need to address this issue and include some appropriate measures in the evaluation process.

5. ACKNOWLEDGEMENTS

I would like to thank my supervisor Udo Kruschwitz for his help, advice and helpful feedback on drafting this paper, and two anonymous reviewers for their helpful feedback on an earlier version of this document.

References

Agichtein, E., Brill, E., Dumais, S. T. and Ragno, R. (2006), ‘Learning user interaction models for predicting web search result preferences’, *Personal Information Management* pp. 251–259.

Albayrak, S., Wollny, S., Varone, N., Lommatzsch, A. and Milosevic, D. (2005), ‘Agent technology for personalized information filtering: The pia-system’, *ACM Symposium on Applied Computing* **1**(1), 54–59.

Baraglia, R. and Silvestri, F. (2007), ‘Dynamic personalization of web sites without user intervention’, *Communications of The ACM* **50**(2), 63–67.

Bayir, M. A., Toroslu, I. H., Cosar, A. and Fidan, G. (2009), Smart miner: A new framework for mining large scale web usage data, *in* ‘Proceedings of WWW 2009’, ACM, pp. 161–170.

Corsini, P. and Marcelloni, F. (2006), ‘A fuzzy system for profiling web portals users from web access log’, *Journal of Intelligent and Fuzzy Systems* **17**(1), 503–516.

Cutrell, E., Robbins, D. C., Dumais, S. T. and Sarin, R. (2006), ‘Fast, flexible filtering with phlat - personal search and organization made easy’, *Personal Information Management* **1**, 261–269.

Eirinaki, M. and Vazirgiannis, M. (2003), ‘Web mining for web personalization’, *ACM Transactions on Internet Technology* **3**(1), 1–27.

France, T., Yen, D., Wang, J.-C. and Chang, C. M. (2002), ‘Integrating search engines with data mining for customer-oriented information search’, *Information Management and Computer Security* **10**(5), 242–254.

Girardi, R. and Marinho, L. B. (2007), ‘A domain model of web recommender systems based on usage mining and collaborative filtering’, *Requirements Eng* **12**(1), 23–40.

Hu, J., Wang, G., Lochovsky, F., Sun, J.-T. and Chen, Z. (2009), Understanding users query intent with wikipedia, *in* ‘Proceedings of WWW 2009’, ACM, pp. 471–480.

Kosala, R. and Blockeel, H. (2000), ‘Web mining research: a survey’, *SIGKDD Explorations* **2**(1), 1–15.

Kruschwitz, U., Webb, N. and Sutcliffe, R. (2008), *Query Log Analysis for Adaptive Dialogue-Driven Search*, Query Log Analysis, Information Science Reference, 389-414, chapter XX.

Mobasher, B., Cooley, R. and Srivastava, J. (2000), ‘Automatic personalization based on web usage mining’, *Commmunications of the ACM* **43**(8), 142–150.

Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R. (2008), ‘A web usage mining framework for mining evolving user profiles in dynamic web sites’, *IEEE Transactions on Knowledge and Data Engineering* **20**(2), 202–215.

Pazzani, M. J. and Billsus, D. (2007), *Content-Based Recommendation Systems*, Vol. 4321 of *Adaptation Technologies*, Springer-Verlag Berlin Heidelberg 2007, 325-341, chapter II.

Perkowitz, M. and Etzioni, O. (1997), ‘Adaptive web sites: an ai challenge’, *Artificial Intelligence* **11**(1), 246–271.

Perkowitz, M. and Etzioni, O. (1999), ‘Adaptive web sites: Conceptual cluster mining’, *Artificial Intelligence* **17**(1), 243–273.

Perkowitz, M. and Etzioni, O. (2000), ‘Towards adaptive web sites: Conceptual framework and case study’, *Artificial Intelligence* **118**(1), 245–275.

Schafer, J., Frankowski, D., Herlocker, J. and Sen, S. (2007), *Collaborative Filtering Recommender Systems*, Vol. 4321 of *Adaptation Technologies*, Springer-Verlag Berlin Heidelberg 2007, 291-324, chapter II.

Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.-N. (2000), ‘Web usage mining: Discovery and applications of usage patterns from web data’, *SIGKDD Explorations* **1**(2), 12–23.

Teevan, J., Dumais, S. T. and Horvitz, E. (2005), ‘Beyond the commons: Investigating the value of personalizing web search’, *User Modeling and User-Adapted Interaction* **13**(1), 311–372.

Query Performance Prediction Based on Ranking List Dispersion

Joaquín Pérez-Iglesias
Calle Juan del Rosal, 16. 28040 Madrid. Spain.
Universidad Nacional de Educación a Distancia.
http://nlp.uned.es/ jperezi
joaquin.perez@lsi.uned.es

Abstract

In this paper we introduce a novel approach for query performance prediction based on ranking list scores dispersion. Starting from the premise that different score distributions appear for good and poor performance queries, we introduce a set of measures that capture these differences between both types of distributions. The proposed measures will employ the ranking list, output of a search system, as an information source to predict query performance in terms of MAP. The obtained results reveal a significant correlation degree with MAP and are very similar to those achieved with more complex methods. Finally some generic open questions that could guide further research on query prediction methods are introduced.

Keywords: Query Performance Prediction

1. INTRODUCTION

During the last years a growing attention has been focused on the problem of query performance prediction. This topic has turned into an important challenge for the IR community. Query performance prediction deals with the problem of detecting those queries for which a search system would be able to return a document set useful for an user. In other words query prediction objective is the development of a search system able to estimate the quality of its answer before the relevant document set is supplied to the user.

A wide range of possible applications appears for a system like the described before. For example a system could ask the user for some extra information in order to improve the result quality before an answer is supplied; a federated search system could select the best answer from any of its sources; a specialised search system on one specific subject could decide by itself if it needs make use of a broader topic index in order to supply a better answer.

In this paper a novel approach for query performance prediction is introduced. The proposed method falls into post-retrieval prediction methods. This type of predictors make use of the information supplied from the search system, once the search has been carried out, opposite to pre-retrieval prediction where the estimation is computed before the search has been completed. The proposed method is based on the study of document scores distribution among the document scores ranking list. This proposal is based on the hypothesis that some differences between the scores distribution of good performance and poor performance queries can be observed. Some measures that try to capture the prior differences among document scores from a ranking list will be proposed in order to predict query performance.

1.1. Related work

In the last years several works dealing with query performance prediction have been proposed. In general the different prediction methods can be classified into two main groups, those approaches that use information from the results obtained after a query is executed (post-retrieval predictors), and those that try to estimate query difficult before the ranking list is obtained from the search engine (pre-retrieval predictors). In general it is accepted that the last has as main advantage a low computational cost even at the expense of providing less accurate estimations.

Pre-retrieval predictors use statistics as collection frequency (CF), inverse document frequency (IDF) or query length. These methods try to detect the ambiguity of the query based on these statistics. He and Ounis [9] propose different measures based on IDF and CF as the inverse collection term frequency mean or the IDF standard deviation. Recently Zhao et al. [17] have proposed some measures based on the standard deviation of the query terms, which are weighted using a TF-IDF schema.

Post-retrieval methods are more related to the approach introduced in this paper. First attempts were started by [5] where *Clarity Score* was proposed. This estimator tries to measure the ambiguity of a query with respect to the document collection. The ambiguity of a topic is calculated with Kullback-Leibler divergence (KLD) between the collection and the top ranked documents language model. A good performance query will show a high divergence value, it can be explained by the fact that top ranked documents are about a single topic and this involves low ambiguity. Further works based on Clarity Score like *Ranked List Clarity Score*, which replace ranking scores by ranking position, and *Weighted Clarity Score* that assigns different weights to query terms in order to calculate KLD, appears in [6].

More specific methods for the web environment can be found in [18], *Weighted Information Gain* and *Query Feedback*, where both measures show a good performance for ‘ad-hoc’ and Named Pages tasks. The first method tries to measure the information gain between a state where an average document is retrieved and the state where the real search has been accomplished. On the other hand *Query Feedback* models the retrieval system as a noisy channel. The input for the noisy channel is the query and the output is the ranking list obtained. From this premise the authors try to measure the degree of noise introduced by the retrieval system and from it to estimate the query performance.

Carmel et al. [4], try to model the relation among the three components that take part into the process of retrieval: topic, set of relevant documents and the collection. The relation between them is measured by means of the Jensen-Shannon divergence. Once the previous measures have been calculated they propose the application of a machine learning method to combine them.

In the works developed by Yom-Tov et. al [16] the proposed model was based on the agreement between the full query and sub-queries, where each one of these sub-queries include only one term from the original query. ‘Agreement’ is measured from the overlap between full query and sub-queries ranking list results. They conclude that in general hard queries will not show agreement between the obtained ranking list, while a high level of overlapping will mean a good performance query.

Aslam and Pavlu [3] propose a technique based on the Jensen-Shannon divergence between the ranking lists obtained from multiple retrieval functions. This approach is based on the idea that for ‘easy’ queries ranking functions must agree and therefore a lower divergence would be calculated.

A different approach based on KLD was proposed by Amati et al. [1]. Here the term frequency divergence between top retrieved documents and the whole collection is measured. They claim that a well-defined query (good performance topic), will show a significant divergence.

Recently a new improved version of clarity score has been presented by Hauff et. al [8]. The authors propose two main contributions to clarity score. First, the set of the feedback documents used in order to compute the query language model is fixed to the documents that contain all query terms. Next, they propose to select a subset of terms from the retrieved documents in order to remove the noise generated by those terms with a high document frequency.

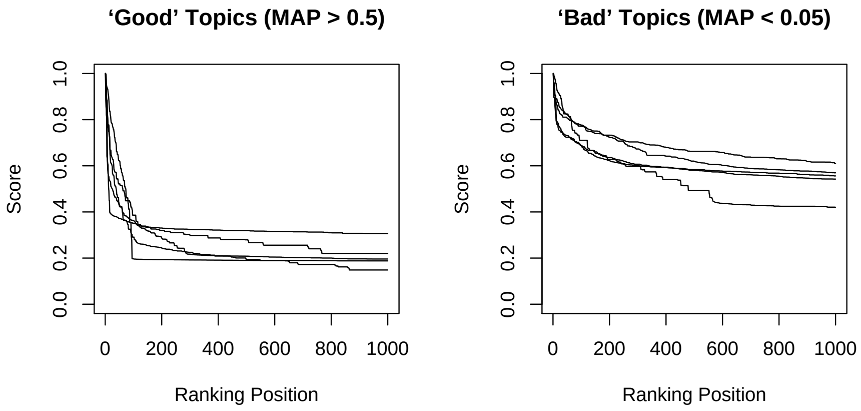


Figure 1: 5 Best Performing Topics (left) Vs 5 Worst Performing Topics (right), from Robust 2004 using BM25. Scores have been normalised in $[0, 1]$. The maximum number of retrieved documents has been fixed to 1000.

Following, a set of new post-retrieval predictors based on the scores dispersion among a ranking list will be introduced. These measures will employ the ranking list obtained from a retrieval system in order to predict the query performance, with a low computational cost.

2. RANKING LIST SCORES DISPERSION AS A PREDICTOR

The approach proposed on this paper is based on the study of the ranking list obtained after a retrieval process is executed. As it is well known a ranking function tries to order documents based on their relevance for a topic. For this purpose a ranking function assigns a weight (or score) to each document in the collection. In a naive sense this score can be interpreted as a 'quantitative measure' of the document relevance. The ranking list scores distribution can be an indicative of the quality performance for a specific topic. Based on this premise some differences between document scores distribution for good and poor performing topics should be observed.

For example, if a ranking list has a high value of dispersion among the document scores, it could be a sign that the ranking function has been able to discriminate between relevant and not relevant documents. On the other hand if a low level of dispersion appears, because the ranking function has assigned similar weights, it can be interpreted as it was not able to distinguish between relevant and not relevant documents.

The differences in terms of scores dispersion can be observed in figure 1, where some of the best and worst performance topics for Robust 2004 are represented. As it can be seen best topics show a longer distance between maximum and minimum score and a sharp slope. Opposite to this, topics with poor performance show a lower distance between maximum and minimum and a softer slop.

A feature of this approach is that it can be applied independently of the ranking model employed for retrieval. The reason for this characteristic is due to retrieval models try to maximise in terms of score, the differences between relevant and not relevant documents[1]. Based on this property the proposed approach can be considered as a generic method of quality performance prediction, not dependent on the model applied for document weighting.

The measures that will be introduced on next section are focused on trying to capture the differences between good and poor performance topics in terms of dispersion. In order to evaluate the quality of the proposed measures the correlation between them and AP (*average precision*) will be computed.

[1] At least for probability based ranking models.

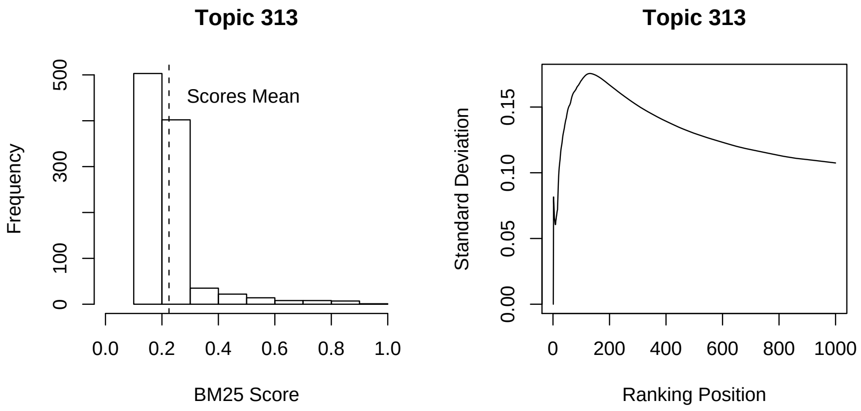


Figure 2: Scores histogram and scores standard deviation respectively for Topic 313. Scores have been normalised in range $[0, 1]$. The maximum number of retrieved documents has been fixed to 1000.

2.1. Proposed Measures

A reliable method to capture and measure dispersion along the obtained ranking list is a key part of this work. Some prior studies have tried to model how document weights are distributed along a ranking list. In general, and as a simplification, it can be assumed that an adequate model could be a mix between an exponential and a normal probability distribution. Exponential for not relevant documents, and normal for relevant documents [12, 13]. Generally a great number of retrieved documents are not relevant (exponential distribution), thus it is likely that a great number of documents will be weighted with a low score. A consequence of it is that ranking lists shape use to hold a long *tail* where a majority of not relevant documents are placed, see figure 1. It is important to understand how the scores distribution will affect some of the typical measures in order to capture the dispersion.

Some notation is needed to define the following proposed measures: (i) A ranking list RL is a document list sorted in decreasing order by their documents scores; (ii) The score assigned to a document d, placed at position i into the ranking list, by a ranking function is defined as $score(d_i)$.

1. **Minimum normalised score:** A first approach to measure the dispersion can be computed with the minor document score found into the ranking list. The minimum score must be normalised in order to make possible a comparison among it and the minimum scores obtained by the rest of topics. Thus for a ranking list of size N:

$$Min = \frac{score(d_N)}{score(d_1)}$$

2. **Standard Deviation:** Standard deviation σ is a simple measure of the variability or dispersion of a data set. A low standard deviation indicates that the data points tend to be very close to the same value (the mean μ), while high standard deviation indicates that the data are spread out over a large range of values. Given ranking list scores mean $\mu(RL)$, standard deviation is computed as next:

$$\sigma(RL) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(score(d_i) - \mu(RL))^2}$$

A drawback in the use of the standard deviation is caused by the great number of low scores assigned by the ranking function. As was described previously, a high percentage of document scores have a low value, which causes that mean is displaced towards the region of densest distribution, that is the tail of the ranking list as can be seen in figure 2 (left). As a consequence of it, the deviation on the top documents is not captured properly when the standard deviation is computed along the full ranking list.

3. **Maximum Standard Deviation:** A different approach to minimise the effect of low scores high frequency is computing the maximum standard deviation. This estimator is based on the idea of computing the standard deviation at each point in the ranking list, and selecting the maximum standard deviation found. As can be seen in figure 2 (right), which shows how standard deviation evolves along ranking list, this measure tends to decrease once the maximum has been reached, coinciding with the start of the *ranking list tail*. The σ_{max} is defined as next:

$$\sigma_{max} = max[\forall d \in RL, \sigma(RL_{[1,d]}))]$$

Following, the experimental setup designed in order to test the validity of the proposed measures will be described.

3. EXPERIMENTAL SETUP

The validity of a query performance predictor is tested based on the correlation coefficient between the proposed predictor and the performance of the real search system in terms of AP. Besides correlation coefficients, a standard test collection and a set of state-of-art retrieval models have been used.

Correlation: On related literature three different correlations coefficients can be found: *Pearson*, *Spearman*(ρ) and *Kendall*(τ). *Pearson* indicates the strength and direction of a linear relationship between two data series. *Kendall* and *Spearman* are based on ranking correlation between both data series. In both cases values are ranked and the correlation coefficient depends on the observable differences between both ranks. More specifically *Spearman* applies a basic linear correlation between both rankings while *Kendall* computes the correlation value counting pairwise swapping needed in order to transform one ranking into the other. The three correlation coefficients calculate a real number in the range $[-1, 1]$, where 1 means perfect correlation, -1 means a perfect inverse correlation and 0 means no correlation at all.

Test data: The different measures proposed in this paper have been tested with the set of documents from TREC Disk4 & 5, minus Congressional Record. This data was employed for the Robust Track 2004 and contains around 528,000 documents with a total size of almost 2 GB [14]. The set of topics are those available from the same track, that is topics 301-450 and 601-700[2], this makes a total of 249 topics with their relevant judgement. Only the field title from topics has been employed in the experiments.

Ranking models: Since the proposed method can be applied to any retrieval model, we have selected a set of retrieval models representative enough to test the validity and compare the obtained predictions values among them. For each retrieval model[3] the parameters have been fixed to the values recommended by Terrier documentation:

- Okapi BM25 [11] with parameters $b = 0.34$, $k_1 = 1.2$ and $k_3 = 8$.
- Language Model proposed by Hiemstra in [10] with $\lambda = 0.15$.
- DFR-PL2 proposed by Amati et. al in [2] with $c = 9.150$.

4. RESULTS

The obtained results[4] after the execution of the experiments are shown at table 1. These measures were computed with a default ranking list size of 1000. This is the standard size in TREC experiments for the MAP calculation. As it can be seen the obtained correlation coefficients, among the different retrieval models, are similar. This similarity, in the obtained results with the

[2] Topic 672 has been removed since no relevant documents can be found for it in the collection.
[3] The retrieval models have been tested with the Terrier search engine developed by *The University of Glasgow*.
[4] Obtained results are statistically significant at a level of 0.01.

	BM25			LM			PL2		
	Sp	Pe	Ke	Sp	Pe	Ke	Sp	Pe	Ke
Min	-0.53	-0.49	-0.37	-0.54	-0.42	-0.38	-0.53	-0.45	-0.37
σ	0.49	0.39	0.34	0.48	0.35	0.33	0.43	0.30	0.29
σ_{max}	0.57	0.40	0.41	0.54	0.40	0.39	0.52	0.37	0.37

Table 1: Spearman(Sp), Pearson(Pe) and Kendall (Ke) correlation coefficients obtained with the proposed measures.

different retrieval models, can be interpreted as a proof of a common behaviour in terms of scoring, as was suggested before.

Minimum score measure shows a significant correlation degree, although its performance could drop when a whole ranking list is used. This is a consequence of the ranking list trend to achieve zero at the lower positions, when the scores have been normalised by the maximum score found. As was expected and due of the *ranking list tail* effect described before, standard deviation shows a poor performance measuring the scores dispersion.

On the other hand the results obtained with the maximum standard deviation outperforms to those achieved with standard deviation. Therefore σ_{max} avoids, at least in part, the lack of precision, in terms of dispersion measurement, obtained by the classic standard deviation.

5. CONCLUSIONS

In this paper some novel query performance predictors have been introduced. The obtained results show that measures based on standard deviation over scores ranking list, can be used to predict the quality of a search system reply.

The application of standard deviation as a dispersion measure for a ranking list, has shown to be a weak approach due to the noise introduced by the not relevant documents set retrieved. In order to avoid the prior effect the σ_{max} has been applied and it has improved the results acting as a noise reduction method. The correlation degree has been calculated with the most important correlation coefficients obtaining for all of them similar results. The obtained results outperform pre-retrieval approaches with a similar computational cost. In relation with post-retrieval approaches the results obtained are similar with the advantage of a much lower computational cost.

The consistency of the obtained results with the different retrieval models suggests the validity of the proposed method independently of the retrieval model. The main reason of this generalisation is achieved due it is based on a common expected behaviour of any ranking retrieval model, that is, the ability to distinguish between relevant and not-relevant documents for a specific topic.

6. FUTURE WORK

As a consequence of the study of the related work and the development of a new family of predictors, some open questions have been identified which could guide further research about this topic. Firstly it has not been established clearly which correlation coefficient is more adequate. Evaluations based on any of the three traditional correlation coefficients can be found, these are applied indistinctly without a clear justification of its use. This issue has been recently discussed by Claudia Hauff et al. in [7].

Moreover an evaluation based only on correlation coefficients can be hard to interpret. As can be found on the related literature, the obtained results in terms of correlation degree between prediction methods are almost equivalent for many of them. Some other measures have been proposed as *Kendall Average Precision* by Yilmaz et al. [15], and *Root Mean Square Error* in [7]. In our opinion a new family of different evaluation measures should be proposed, these measures should be more focused on the qualitative aspects of the prediction methods, and thus it should guide us towards a better understanding of the possible applications of prediction methods.

Finally, the application of prediction methods to improve some of the typical tasks related with information retrieval, as query expansion or search systems responses selection, should be further studied. We will try to answer the open questions raised on this work in further research.

Bibliography

[1] Giambattista Amati, Claudio Carpineto, and Giovanni Romano. Query difficulty, robustness, and selective application of query expansion. In *ECIR*, pages 127–137, 2004.

[2] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.

[3] Javed A. Aslam and Virgil Pavlu. Query hardness estimation using jensen-shannon divergence among multiple scoring functions. *Advances in Information Retrieval*, 4425:198–209, 2007.

[4] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. What makes a query difficult? In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 390, New York, 2006. ACM Press.

[5] Steve Cronen-Townsend, Yun Zhou, and W. Bruce. Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02*, New York, 2002. ACM Press.

[6] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Precision prediction based on ranked list coherence. *Information Retrieval.*, 9(6):723–755, 2006.

[7] Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. *The Combination and Evaluation of Query Performance Prediction Methods*. 2009.

[8] Claudia Hauff, Vanessa Murdock, and Ricardo Baeza-Yates. Improved query difficulty prediction for the web. *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, page 439, 2008.

[9] Ben He and Iadh Ounis. Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval*, pages 43–54, 2004.

[10] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede, January 2001.

[11] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 36(6):779–808, 2000.

[12] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01*, pages 267–275, 2001.

[13] Stephen Robertson. On score distributions and relevance. *Advances in Information Retrieval*, 4425:40–51, 2007.

[14] Ellen M. Voorhees. Overview of the trec 2004 robust retrieval track. In *In Proceedings of the Thirteenth Text REtrieval Conference (TREC)*, 2004.

[15] Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594, New York, NY, USA, 2008. ACM.

[16] Elad Yom-Tov, Shai Fine, David Carmel, and Adam Darlow. Learning to estimate query difficulty. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, page 512, 2005.

[17] Ying Zhao, Falk Scholer, and Yohannes Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. *Advances in Information Retrieval*, 4956:52–64, 2008.

[18] Yun Zhou and W. Bruce Croft. Query performance prediction in web search environments. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, page 543, 2007.

SIREn: Entity Retrieval System for the Web of Data

Renaud Delbru
Digital Enterprise Research Institute
National University of Ireland
Galway, Ireland
renaud.delbru@deri.org

Abstract

We present ongoing work on the Semantic Information Retrieval Engine (SIREn), an "entity retrieval system" specifically designed to meet the requirements of indexing and searching a large amount of semi-structured data, e.g. the entire Web of Data. SIREn supports efficient full text search with semi-structural queries and exhibits a concise index, constant time updates and inherits Information Retrieval features such as top-k queries, efficient caching and scalability via distribution over shards. We demonstrate how SIREn can effectively answer queries over 10 billion triples on single commodity machine. The prototype is currently in use in the Sindice search engine which index at the present time more than 50 million harvested documents containing semi-structured data.

Keywords: Web of Data, Resource Description Framework, Semi-Structured Data, Inverted Index, Scalability

1. INTRODUCTION

The amount of Resource Description Framework[1] (RDF) and Microformats available online has grown tremendously in the past years. RDF and Microformats are specifications that allow web sites to expose semi-structured information for machine reuse, implementing something referred to as the "Web of Data", Semantic Web [13] or even Web 3.0. In a way, the idea behind the Web of Data is to "create a universal medium for the exchange of data where data can be shared and processed by automated tools as well as by people"[2]. Typical uses of Web of Data are automatic interactions with advanced clients, e.g. automatic integration with the user's calendar and contact list by encoding entities using formats such as HCard, FOAF, or VCard, search engine result customization, advanced data mashups, etc.

On the one hand the Linked Open Data[3] community has made available several billion triples equivalent of information, driven by the idea of open access to semi-structured data. On the other hand, an increasing number of relevant Web 2.0 players (LinkedIn, Yahoo Locals, Eventful, Digg, Youtube and Wordpress to name just a few) have also added some form of semi-structured data markups given the support now available in Yahoo with the Searchmonkey project, and in Google with its support for RDFa structured snippets.

Precise measurements of this growth are not available, but partial reports and private communications of which we are aware, estimate it in several billions of RDF triples and approximately half a billion of marked up web pages, likely totalling several billion triples equivalent. Whatever the current size of the Web of Data is today, the trend is clear and so is the requirement for handling semantically structured data with a scalability in the same class of traditional search engines.

In this paper we present the Semantic Information Retrieval Engine, SIREn, a system based on Information Retrieval (IR) techniques designed to search "entities" and exhibiting many characteristics of IR systems such as web like scalability, incremental updates, top-k queries and efficient caching among others.

[1] Resource Description Framework: `http://www.w3.org/RDF/`
[2] Semantic Web Activity Statement: http://www.w3.org/2001/sw/Activity.html
[3] Linked Data: `http://linkeddata.org/`

1.1. Web of Data: Preliminaries

The Web of Data is based on the RDF data model that provides the functionality of making machine understandable statement about any resources. An RDF statement is expressed as a triple (s, p, o) consisting of a subject, a predicate, and an object and asserts that a subject has a property (predicate) with some value (object). Given three infinite sets U, B and L called respectively URI references, blank nodes and literals, an RDF statement (s, p, o) is an element of $(U \cup B) \times U \times (U \cup B \cup L)$. An RDF statement can be interpreted as a labelled edge where both the subject and object are the nodes and the predicate is a labelled edge connecting the two nodes. While RDF triples are a seemingly simple concept, the true power of RDF lies in the fact that these triples are combined to form a labelled, directed multi-graph, as depicted in Fig. 1a.

The Web of Data is composed of many interconnected RDF graphs, or datasets, each one nameable by an URI. These named graphs [5] are composed of quads. A quad q is a statement (s, p, o, c) with a fourth element $c \in U$ called "context" for naming the RDF dataset in order to keep the provenance of the RDF data.

1.2. Web of Data: Requirements for SIREn

SIREn has been conceived to index the entire "Web of Data". The requirements have therefore been:

1. Support for the multiple formats which are used on the Web of Data;
2. Support for entity centric search;
3. Support for context (provenance) of information: entity descriptions are given in the context of a website or dataset;
4. Support for semi-structural full text search, top-k query, scalability via shard over clusters of commodity machines, efficient caching strategy and real-time dynamic index maintenance.

With respect to point 1, the two formats which enable the annotations of entities on web pages are Microformats and RDF. At knowledge representation level, the main difference between Microformats and RDF is that the former can be seen as a frame model while the latter has a graph based data model. While these are major conceptual differences, it is easy to see that the RDF model can be used effectively to map Microformats[4]. Under these conditions, we have developed SIREn to cover the RDF model knowing that this would cover Microformats and likely other forms of web metadata.

With respect to point 2 and 3, the main use case for which SIREn is developed is entity search: given a description of an entity, i.e. a star-shaped queries such as the one in Fig. 1b, locate the most suitable entities and datasets. This means that, in terms of granularity, the search needs to move from "page" (as per normal web search) to a "dataset-entity". The Fig. 1a shows an RDF graph and how it can be split into three entities *renaud*, *giovanni* and *DERI*. Each entity description forms a sub-graph containing the incoming and outgoing relations of the entity node.

Finally, we will see in Sect. 2 that the SIREn model enables dataset-entity centric search while leveraging well known IR techniques to address the point 4.

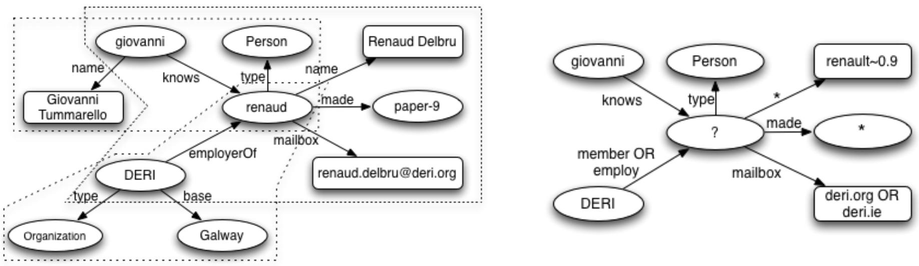
1.3. Approaches for Entity Retrieval

Given an query, an Entity Retrieval System (ERS) helps to locate and retrieve a list of relevant entities. An ERS should allow "imprecise" or "fuzzy" queries and rank the results based on their relevance to the queries. The entity retrieval task is *selection-oriented*, the aim is to select potential relevant entities (e.g. the top-k most relevant ones) from a large entity collection. Two main approaches have been taken for entity retrieval, DBMS based and IR based.

1.3.1. DBMS based approaches

Typically, entities described in RDF data are handled using systems referred to as "triplestores" or "quadstores" and that usually employ techniques coming from the DBMS world. Some of these are

[4] Any23: `http://code.google.com/p/any23/`



(a) Visual representation of an RDF graph. The RDF graph is divided (dashed lines) into three entities identified by the node *renaud*, *giovanni* and *DERI*

(b) Star-shaped query matching the entity *renaud* where *?* is the bound variable and $\star$ a wildcard

FIGURE 1: In these graphs, oval nodes represent resources and rectangular ones represent literals. For space consideration, URIs have been replaced by their local names.

built on top of existing DBMS such as Virtuoso[5] or column stores [1] while others are purposely built to handle RDF, e.g. [7, 16, 11].

These triplestores are built to manage large amounts of RDF triples or quads and they do so employing multiples indices (generally b-trees) for covering all kind of access patterns of the form `(s,p,o,c)`. As for DBMS, the main goal of these systems is answering possibly complex queries, e.g. those posed using the SPARQL query language[6]. This task is a superset of entity retrieval as we defined it, and can be seen as *transformation-oriented* since DBMS provide functionality to select entities but also to transform the result set (e.g. create a new RDF graph). This comes at the cost of maintaining complex data structures and index duplication. Also they usually do not support "imprecise" and top-k queries, but instead return all results that precisely match the query (similarly to SQL query in relational databases).

1.3.2. Information Retrieval for Semi-Structured Data and RDF

Information Retrieval (IR) techniques [2, 9] offer tools to overcome such limitations and have been shown by modern search engines to scale to the size of the web. For these reasons, recent systems [17, 3] have started to explore IR also for searching RDF data.

SIREn continues on this trend of research by proposing a variant of a word level inverted index based on a node-labelled tree data structure. Such technique is coming from IR for semi-structured text [10] such as XML document. Node labelling schemes [14] have been developed to optimise retrieval of XML search engines since they provide an efficient way to encode and query the tree structure of an XML document. In this paper, we propose to adapt such technique to the RDF data model.

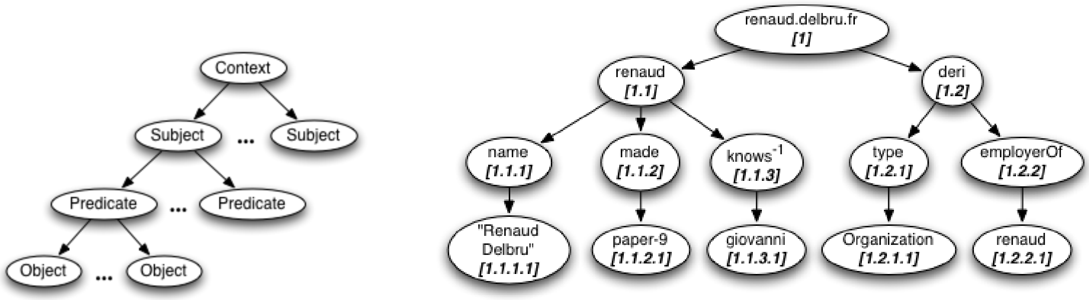
The paper is organized as follows: we first present the index data model in Sect. 2.1 and the associated query model in Sect. 2.2. We report in Sect. 2.3 a short overview of the results of the current scalability evaluation.

2. THE SIREN MODEL

An entity description is a set of triples having one common node and can be depicted as a star-shaped graph. The simplest semi-structured indexing method is to represent an entity as a set of attribute-value pairs using field-based approach [8]. With field-based indexing, index terms are constructed by concatenating the field name with the terms from the content of this field. For example, if a publication has a field *title*, *author* and *abstract*, the index terms for the author field will be represented as *author:renaud* and *author:delbru*.

While this technique is widely used, it has strong limitations when dealing with large amount of heterogeneous semi-structured data:

[5] Virtuoso: `http://virtuoso.openlinksw.com/`
[6] SPARQL: `http://www.w3.org/TR/rdf-sparql-query/`



(a) Conceptual representation of the data tree model of the index

(b) Node-labelled data tree of the example dataset using Dewey's encoding

FIGURE 2: The SIREn data model

1. It is very inefficient to search across all the fields. A single query term will be expanded into a disjunction of n index terms with n the number of different fields.
2. The lexicon becomes prohibitly large, and therefore index term lookups becomes expensive. Many identical terms can appear in various fields, but will be considered as different terms by the system. For example, if m terms appear in n fields, it will produce $m * n$ index terms.
3. Multi-valued fields cannot be handled properly. At query time, we cannot differentiate if an index term belongs to the first or second value of a field. This causes false-positive when using term conjunction. For example, if a publication has two authors, e.g. *Renaud Delbru* and *Giovanni Tumarrello*, then the query *author:renaud AND author:tumarrello* will return a match which is not the expected behaviour.

In order to overcome these limitations, we designed an inverted index based on a tree-structured data model that enables versatile star-shaped querying while enjoying efficient use of disk space, effective compression, fast dynamic updates and sub-linear query processing.

2.1. SIREn Data Model

SIREn, similarly to XML information retrieval engine, adopts a tree data structure and orderings of tree nodes to model datasets, entities and their RDF descriptions. The data tree model is pictured in Fig. 2a. This model has a hierarchical structure with four different kind of nodes: context (dataset), subject (entity), predicate and object. Each node can refer to one or more terms. In case of RDF, a term is not necessarily a word (as in part of an RDF Literal), but can be an URI or a local blank node identifier.

Inverted index based on tree data structure enables to efficiently establish relationships between tree nodes. There are two main types of relations: *Parent-Child* (PC) and *Ancestor-Descendant* (AC). To support this set of relations, the requirement is to assign unique identifiers (node labels) that encodes relationships between the tree nodes. Several node labelling schemes have been developed and the reader can refer to [14] for an overview of them. In the rest of the paper, we will use a simple prefix scheme, the *Dewey Order* encoding [4], but the model is not restricted to the Dewey scheme and another scheme (e.g. interval-based) could be used instead.

Using this labelling scheme, structural relationships between elements can be determined efficiently. An element *u* is an ancestor of an element *v* if label(*u*) is a prefix of label(*v*). Fig. 2b presents a data tree where nodes have been labelled using Dewey's encoding. Given the label $\langle 1.2.1.1 \rangle$ for the term `Organisation`, we can efficiently find that its parent is the predicate `rdf:type`, labelled with $\langle 1.2.1 \rangle$.

The data tree structure with PC and AC relations covers the quad relations CSPO (outgoing relations) and COPS (incoming relations). Incoming relations are symbolised by a predicate node with a $^{-1}$ tag in Fig. 2b. The tree data structure is not limited to quad relations, and could in theory be used to encode longer paths such as 2-hop outgoing and incoming relations.

2.2. SIREn Query Model

Since RDF is semi-structured data, we expect three types of queries: 1. full-text search (keyword based), 2. semi-structural queries (complex queries specified in a star-shaped structure), 3. or a combination of the two (where full-text search can be used on any part of the star-shaped query). We present in this section a set of query operators over the content and structure of the data tree that cover the three types of queries.

2.2.1. SIREn Operators

Content operators The content query operators are the only ones that access the content of a node, and are orthogonal to the structure operators. They include extended boolean operations such as boolean operators (intersection, union, difference), proximity operators (phrase, near, before, after, etc.) and fuzzy or wildcard operators.

These operations allow to express complex keyword queries for each node of the tree. Interestingly, it is possibly to apply these operators not only on literals, but also on URIs (subject, predicate and object), if URIs are normalized (i.e. tokenized). For example one could just use an RDF local name, e.g. `name`, to match `foaf:name` ignoring the namespace.

Structure operators In the following, we define a set of operations over the structure of the data tree. Thanks to these operations, we are able to search content to limited tree nodes, to query node relationships and to retrieve paths of nodes matching a given pattern. Joins over paths are possible using set operators, enabling the computation of entities and datasets matching a given star-shaped query.

Ancestor-Descendant: A//D A node A is the ancestor of a node D if it exists a path between A and D. For example, the SPARQL query in Listing 1, line 1, can be interpreted as an Ancestor-Descendant operator, line 2, and will return the path $\langle 1.2.2.1 \rangle$.

Parent-Child: P/C A node P is the parent of a node C if P is an ancestor of C and C is exactly one level above P. For example, the SPARQL query in Listing 1, line 3, can be translated into a Parent-Child operator, line 4, and will return the path $\langle 1.1.1.1 \rangle$.

Set manipulation operators These operators allow to manipulate nodes of the tree (context, subject, predicate and object) as sets, implementing union ($\cup$), difference ($\backslash$) and intersection ($\cap$). For example in Listing 1, the SPARQL query, line 5, can be interpreted as two Parent-Child operators with the intersection operator (AND), line 6.

In addition, operators can be nested to express longer path as shown in Listing 1, line 7 and 9. However, the later is possible only if deeper trees have been indexed, i.e. 2-hop outgoing and incoming relations of an entity.

Listing 1: SPARQL queries and their SIREn interpretation

```
 1  SELECT ?g WHERE { GRAPH ?g { deri ?p renaud }}
 2  deri // renaud
 3  SELECT ?g ?s WHERE { GRAPH ?g { ?s name "Renaud Delbru" }}
 4  name / "Renaud Delbru"
 5  SELECT ?g ?o WHERE { GRAPH ?g { giovanni knows ?o . deri employerOf ?o . }}
 6  knows^-1 / giovanni AND employerOf^-1 / deri
 7  SELECT ?s WHERE { GRAPH <renaud.delbru.fr> { ?s knows renaud }}
 8  renaud.delbru.fr // knows / renaud
 9  SELECT ?g ?s WHERE { GRAPH ?g { ?s employerOf ?o . ?o name "renaud" . }}
10  employerOf // name / "renaud"
```

2.2.2. SPARQL Interpretation

In this section we discuss the extension by which, given the above discussed operators, it is possible to support a subset of the standard SPARQL query language.

By indexing outgoing relations alone, we can show to cover the quad access patterns listed in Table 1. A quad lookup is performed using the tuple operators. Join operations over these patterns

SPOC	POCS	OCSP	CPSO	CSPO	OSPC
(?,*,*,?)	(?,p,*,?) p	(?,*,o,?) o	(?,*,*,c) c	(s,*,*,c) c/s	(s,*,o,?) s//o
(s,*,*,?) s	(?,p,o,?) p/o	(?,*,o,c) c//o	(?,p,*,c) c//p	(s,p,*,c) c/s/p	
(s,p,*,?) s/p	(?,p,o,c) c//p/o	(s,*,o,c) c/s//o			
(s,p,o,?) s/p/o					

TABLE 1: Quad patterns covered by outgoing relations and their interpretation with the SIREn operators. The *?* stands for the elements that are retrieved and the * stands for a wildcard element.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Time (s)	0.75	1.3	1.4	0.5	1.5	1.6	4	0.35
Hits	7552	9344	3.5M	57K	448	8.2M	20.7M	672

TABLE 2: Querying time in seconds and number of hits for the 10 billion triples benchmark

are also feasible. Intersection, union and difference between two or more quad patterns can be achieved efficiently using set manipulations over tree nodes.

The covered quad patterns are a subset of the quad patterns covered by conventional RDF data management systems [6]. In fact, they give the ability to retrieve information about variables that are restricted to be at the subject, object or context position.

It is important to underline however that we are restricting the search of an entity inside a dataset, i.e. SIREn does not allow the use of joins over different contexts, and the intersection of quad patterns within an entity, i.e. SIREn does not allow the use of chains of joins among multiples entities. This limits the query expressiveness to a star-shaped query, e.g. in Fig. 1b.

2.3. Evaluation

We finally report some performance and scalability benchmarks. Indexing time for a synthetic dataset (120GB, 1 billion triples) took only 31 minutes while keeping constant time update and a relatively concise index (15GB) due to efficient compression using word-aligned binary codes. On a term-interleaved inverted index, we achieved a compression average of less than one byte per integer using Simple-9. Query time execution of various star-shaped queries performed at the same order of magnitude than the state-of-the-art triple store RDF-3X [11].

We evaluate SIREn scalability by indexing a dataset composed by 1 billion entities described in approximately 10 billion triples. The dataset is derived from the billion triple challenge dataset[7]. The machine that served for the experiment was equipped with 8GB ram, 2 quad core Intel processors running at 2.23 Ghz, 7200 RPM SATA disks, linux 2.6.24-19, Java Version 1.6.0.06 and GCC 4.2.4.

The following benchmark was performed with cold-cache by using the `/proc/sys/vm/drop_caches` interface to flush the kernel cache and by reloading the application after each query to bypass the application cache. The set of queries is provided at `http://siren.sindice.com`. The performance is given in the Table 2.

3. CONCLUSION

In this paper, we presented the current state of SIREn, an Entity Retrieval System for the Web of Data. The main challenge that has been discussed in this paper is the design of an efficient inverted index especially designed for answering semi-structured (star-shaped) queries while preserving desirable IR features, such as web like scalability, incremental updates, top-k queries and efficient caching among others.

This work has been undergoing for approximately 18 months now and is at different stages of experimentation, validation and ultimately deployment. For example, SIREn presented in this paper

[7] Semantic Web Challenge: `http://challenge.semanticweb.org/`

is currently in use in the Sindice search engine [12] and will be released as an open source project[8]. The system has been evaluated against other existing systems such as RDF-3X, and the results will be published soon. The current work on SIREn is going towards the design of a new labelling scheme extending the basic tree model to a directed acyclic graph, hence enabling more flexible semi-structured queries. Due to space constraint, we limited the discussion to the SIREn model and omitted another aspect of the research work: improving the search quality using link analysis on the Web of Data. An early paper [15] about dataset ranking has been published and the complete ranking system has been finalised and qualitatively evaluated. A paper presenting the results will be published soon.

REFERENCES

[1] D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach. Scalable semantic web data management using vertical partitioning. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 411–422. VLDB Endowment, 2007.

[2] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[3] H. Bast, A. Chitea, F. Suchanek, and I. Weber. ESTER: efficient search on text, entities, and relations. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference*, pages 671–678, New York, NY, USA, 2007. ACM.

[4] K. Beyer, S. D. Viglas, I. Tatarinov, J. Shanmugasundaram, E. Shekita, and C. Zhang. Storing and querying ordered xml using a relational database system. In *SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on Management of Data*, pages 204–215, New York, NY, USA, 2002. ACM.

[5] J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Named graphs, provenance and trust. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2005. ACM.

[6] A. Harth and S. Decker. Optimized index structures for querying rdf from the web. In *LA-WEB*, pages 71–80, 2005.

[7] A. Harth, J. Umbrich, A. Hogan, and S. Decker. YARS2: A Federated Repository for Querying Graph Structured Data from the Web. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, volume 4825 of *Lecture Notes in Computer Science*, pages 211–224. Springer Verlag, November 2007.

[8] R. W. Luk. A survey in indexing and searching XML documents. *Journal of the American Society for Information Science and Technology*, 53(6):415, 2002.

[9] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[10] G. Navarro and R. Baeza-Yates. A language for queries on structure and contents of textual databases. In *SIGIR '95: Proceedings of the 18th international conference on Research and Development in Information Retrieval*, page 93, New York, NY, USA, 1995. ACM.

[11] T. Neumann and G. Weikum. RDF-3X - a RISC-style Engine for RDF. *Proceedings of the VLDB Endowment*, 1(1):647–659, 2008.

[12] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: A document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics and Ontologies*, 3(1), 2008.

[13] N. Shadbolt, T. Berners-Lee, and W. Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006.

[14] H. Su-Cheng and L. Chien-Sing. Node Labeling Schemes in XML Query Optimization: A Survey and Trends. *IETE Technical Review*, 26(2):88, 2009.

[15] N. Toupikov, J. Umbrich, R. Delbru, M. Hausenblas, and G. Tummarello. DING! Dataset Ranking using Formal Descriptions. In *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*, Madrid, Spain, 2009.

[16] C. Weiss, P. Karras, and A. Bernstein. Hexastore - sextuple indexing for semantic web data management. *Proceedings of the VLDB Endowment*, 1(1):1008–1019, 2008.

[17] L. Zhang, Q. Liu, J. Zhang, H. Wang, Y. Pan, and Y. Yu. Semplore: An IR Approach to Scalable Hybrid Query of Semantic Web Data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, volume 4825 of *Lecture Notes in Computer Science*, pages 652–665. Springer Verlag, November 2007.

[8] SIREn: `http://siren.sindice.com/`

Context Features and their use in Information Retrieval

Carla Teixeira Lopes
Doctoral Program in Informatics Engineering of
Faculdade de Engenharia da Universidade do Porto
carla.lopes@fe.up.pt

Abstract

The context in which a search takes place affects the Information Retrieval (IR) process. It affects the searcher's interaction with the IR system, his expectations and his decisions about the documents he retrieves. Therefore, knowing more about what features are important in a searcher's context and what they are used for, can help design more useful and successful IR systems. This paper has three main contributions. It starts with a literature review on the definition of context and on context taxonomies (1). A systematic representation of context features and uses, based on related work, is then proposed (2) and used in a survey on the use of context features in IR (3). This analysis has concluded that *interaction context* is the most used category of features and *Indexing and Searching* are the tasks where context features are most employed. This work, an initial phase of a PhD research, provides a systematic review of what is being done in the area and proposes a taxonomy for IR.

Keywords: Information Retrieval, Context Features, Context Uses

1. INTRODUCTION

Typically, Information Retrieval (IR) systems support their decisions solely on the query and document collection. Several implicit factors about the user and the search context (e.g. time, location, task, expertise, interaction) are ignored and could be considered to optimize IR performance. In fact, all information activities take place within a context that affects the way people access information, interact with a retrieval system, evaluate and make decisions about the retrieved documents (Ingwersen & Järvelin 2005, Harper & Kelly 2006). A contextualised strategy might allow IR systems to learn and predict what information a searcher needs, learn how and when information should be displayed, present results relating them to previous information and to the tasks the user has been engaged in and decide who else should get the new information.

In the field of Information Retrieval, there is a growing interest in improving the search process towards the user needs and context (Bierig & Göker 2006). An early model that has approached IR from the level of context is the one from Belkin (1980). Later, other authors have also developed models (Ingwersen 1996, Saracevic 1997) in which context is at the center of the IR process. Still in the decade of 1980, another project (Saracevic et al. 1988, Saracevic & Kantor 1988a, Saracevic & Kantor 1988b) was dedicated to the characterization of the elements involved in information seeking and retrieving, such as the cognitive context involved in these processes. More recently, several journals (e.g: Information Processing and Management - 2002, 2008; Information Retrieval - 2007) and conferences have given attention to this topic (e.g: Information Retrieval in Context (IRiX) - 2004, 2005; Information Interaction in Context (IIiX) - 2006, 2008).

While there is consensus that context matters (Cool & Spink 2002, Bierig & Göker 2006), there is no agreement on which context elements influence IR (Cool & Spink 2002). The purpose of this paper is threefold. It starts with a literature review on the definition of context and on proposed context taxonomies (Section 2). A systematic representation of context features and uses, based on related work, is then proposed (Section 3) and used in a survey on the use of context features in contextual IR (Section 3.2). Section 5 outlines the PhD work that will be built on the current survey and identifies issues for discussion in the Symposium.

2. BACKGROUND AND RELATED WORK

Context is one of the most abused terms in IR, being associated to a large range of ideas (Finkelstein et al. 2002). Brézillon (1999) enumerates twelve different definitions from several authors where the lack of consensus is evident. As Dervin (1997) says, "context has the potential of being virtually anything ... [it is] a kind of container in which the phenomenon resides". The concept crosses several areas of knowledge from cognitive sciences to engineering. This section reports on definitions in domains related to IR and does not intent to do a thorough review of definitions in other areas. The work of Brézillon (1999) presents a more thorough review of context's definitions in five areas connected to artificial intelligence.

In the literature, some authors have gone further in the characterization of context, defining contextual taxonomies. These structures facilitate the understanding and exploration of context. Some of the main context taxonomies will also be described in this section.

2.1. Context Definition

According to Dourish (2004), *context* may be defined in two perspectives: as a representational problem or as an interactional problem. In the first perspective, it is viewed as a form of information that is delineable, stable and independent of the activity. It consists of implicit attributes that describe the user and the environment in which information activities occur. The second perspective sees context as arising from the activity, from which it can't be separated.

Dey & Abowd (2000) also do an extensive review on context's definitions. They propose their own definition that encompasses other authors' definitions. Context is: "any information that can be used to characterize the situation of entities (e.g. a person, a place or an object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves". This definition matches the first perspective of Dourish. Other authors give definitions of context that also correspond to the first perspective of Dourish. Göker & Myrhaug (2002) present a short and comprehensive definition: "description of aspects of a situation", similar to the one from Dey & Abowd (2000). Marchionini (1997) had already defined it as a "setting" that has "physical and conceptual/social components, including whether the task is done in collaboration or alone and the information seekers physical and psychological states". The first definition proposed by Johnson (2003) also equals context to "situation".

The second definition of Johnson (2003) goes beyond the enumeration of factors to the specification of the active ingredients in a context, noting that they have predictable effects on processes. In this view, context is defined as a relation between the specific ingredients and the processes, which is closer to the second perspective of Dourish. Similarly, Winograd (2001) says that "something is context because of the way it is used in interpretation". Sato (2001) defines context as "a pattern of behavior or relations among variables that are outside of the subjects of design manipulation and potentially affect user behavior and system performance". Ingwersen & Järvelin (2005) say that "actors and their components function act as context to one another in the interaction processes. There are social, organizational, cultural as well as systemic contexts, which evolve over time".

2.2. Context Taxonomies

Ingwersen & Järvelin (2005) present a nested model of context stratification for IR with six dimensions. **Intra-object structures** refers to context obtained from each document where images are contextual to a surrounding text, paragraphs act as context for their own lines and words. **Inter-object contexts** are concerned with the properties of documents, like references, citations, outlinks and inlinks, that give and take context from other objects. **Interaction/session context** is about the social interaction and interactive IR activities, if the searcher is at the core, or is about the retrieval session, if the interface is at the core of the taxonomy. **Social, systemic, media, work task, conceptual, emotional contexts** are related to socio-organizational and systemic aspects (like the IT, interface and documents), if the searcher is at the core, or are related to information objects and searching actors, if the interface is at the core. **Economic techno-physical and societal contexts** correspond to the prevailing societal infrastructures. Finally, **historic context** is a temporal form of context that includes all past participating actors' experience.

Dey & Abowd (2000) propose a classification of context information based on the entities in which the context is assessed and on categories of context. They define three entities: **places** like regions of geographical space such as rooms or offices, **people** including individual or groups, co-located or distributed and **things** (e.g. physical objects or software components and artifacts like a computer file). Primary and secondary context characterize these entities. Primary context types are: **identity**, **location**, **status/activity** and **time**. These context types may be used to infer additional pieces of context such as the address of a person by her identity. The latter are designated by secondary context types. In their work they also propose categories for uses of context: presentation of information/services to the user, execution of a service and tagging of context to information for later retrieval.

Göker & Myrhaug (2002) present a context taxonomy in which context elements are divided into five main categories. The **task** category is about what the user is doing, his goals, tasks, activities. The **social** one refers to the social aspects of the user, such as information about friends and family or his role. **Personal** context aggregates mental and physical information about the user such as mood, expertise, disabilities. In the **spatio-temporal** category are included attributes like time and location and the **environmental** context is about user surroundings like things, light, people and information accessed by the user.

Briconsouf & Newman (2007) propose a framework to analyse the use of context in health care applications. Their framework has three main axes to characterize context. The **purpose of use of context** presents the three types of context uses proposed by Dey & Abowd (2000). The second axis, **items for context representation**, identifies three main classes to split items of context into: people, environment and activities . The third axis, **organization of context features** proposes other ways to organize context features such as an hierarchical organization that draws from general to local aspects of context, an organization according to the internal and external dimension of context and an organization according to the usefulness of context (relevant or non relevant for the current action).

Mansourian (2008) has also developed a taxonomy for the contextualization of web search with five main categories. The **web user axis** is divided in feelings, thoughts (attitudes and cognitive style), actions (passive vs active users) and competence. The **search tool** and the **search topic** are two other axis of the taxonomy. The fourth axis, **search situation** is divided in place of search, type of search (work-related or everyday life search), immediacy of search and importance of search. The last axis, **information resources** is split in searchability and accessibility, level of provision (publicly available/restricted access) and level of user-friendliness.

From all the reviewed taxonomies, only the one from Ingwersen & Järvelin (2005) has been made for IR. This is the most exhaustive taxonomy, even though it doesn't propose a classification for uses of context. Only the Dey & Abowd (2000) and Briconsouf & Newman's (2007) taxonomies include this categorization. Göker & Myrhaug's (2002) taxonomy is a well known taxonomy in the field of IR.

3. PROPOSAL OF A CONTEXT TAXONOMY FOR IR

Ingwersen & Järvelin's (2005) taxonomy is the most appropriate to our goals. Yet, it does not covers uses of context. Therefore, it is here proposed a context taxonomy for IR composed of two categorizations, one for the context features potentially useful in a IR system (Figure 1) and other for possible uses of these features in a IR system (Figure 2). The context features category is a variant of the Ingwersen & Järvelin's (2005) taxonomy.

In this proposal, context is considered an interactional problem, as defined by Dourish (2004). It is considered that it does not only deal with the environmental features surrounding the user and its activities, but also concerns the interaction in other tasks and situations in similar domains. Context evolves over time and users' context can change each time a new search is made, a new set of results is reviewed or a new document is viewed (Harper & Kelly 2006). Therefore, "it arises from and is sustained by the activity itself" (Dourish 2004).

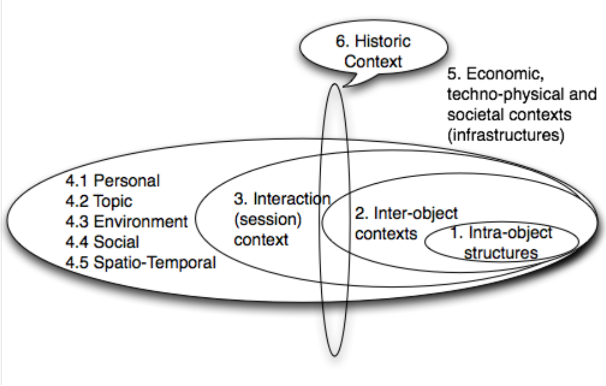

6. Historic Context
5. Economic, techno-physical and societal contexts (infrastructures)
4.1 Personal
4.2 Topic
4.3 Environment
4.4 Social
4.5 Spatio-Temporal
3. Interaction (session) context
2. Inter-object contexts
1. Intra-object structures

FIGURE 1: Taxonomy for Context Features - variant of (Ingwersen & Järvelin 2005)

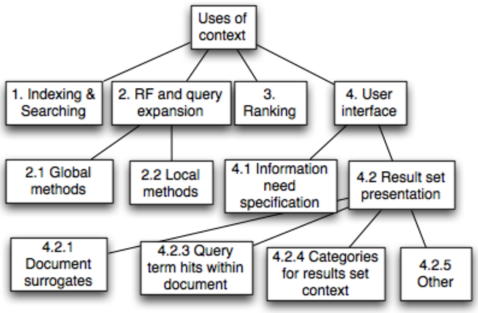

Uses of context
1. Indexing & Searching
2. RF and query expansion
3. Ranking
4. User interface
2.1 Global methods
2.2 Local methods
4.1 Information need specification
4.2 Result set presentation
4.2.1 Document surrogates
4.2.3 Query term hits within document
4.2.4 Categories for results set context
4.2.5 Other

FIGURE 2: Taxonomy for Uses of Context

3.1. Context Features

The proposed taxonomy is similar to the one proposed by (Ingwersen & Järvelin 2005). It dffers in its fourth dimension that is mainly modeled by the context's categories defined by (Göker & Myrhaug 2002). The adoption of the model of Ingwersen & Järvelin is justified by the presence of several relevant dimensions in the IR domain. This model's fourth dimension includes the majority of the other taxonomies' categories. The option for integrating the Göker & Myrhaug's (2002) taxonomy is explained by its comprehensiveness and by its clear and logical partition of context features in categories.

The model is defined by 6 dimensions. Dimensions 1 and 2 are related to the intra/inter-object contexts. The first is related to the documents' structure and content that may act as context. The second is about documents' properties that relate them with other documents. The information resources' category of the Mansourian's (2008) taxonomy fall in the intra-object context category.

In Ingwersen & Järvelin's (2005) model, the third dimension may be approached in two different ways according to who is at the core of the model: the user or the interface. As the proposed taxonomy is centred on the user, this category is about all the social interaction and activities that occur inside the IR session: "what the persons (actors) are doing [...] can be described with explicit goals, tasks, actions, activities, or events. [...] can include other persons tasks (that are within the situation)" as defined by Göker & Myrhaug. Task context can also be characterized by variables like endurance, frequency and stage (Kelly 2006). This dimension contains the activities category in the Briconsouf & Newman's (2007) taxonomy and a part of the status/activity category in the Dey & Abowd's (2000) taxonomy.

The fourth dimension joins the other four categories of the Göker & Myrhaug's (2002) taxonomy. The **personal context** contains the physiological (e.g. "pulse, blood pressure, weight, glucose level, retinal pattern, and hair colour") and the mental context (e.g. "mood, expertise, angriness, and stress"). The **topic context** has information about the persistence and familiarity of the user with the topic (Kelly 2006) and may also contain information about its nature (work or non-work related; fact or subject search) (Mansourian 2008). **Environment context** captures the entities that surrounds the user such as things, services, temperature, light, humidity, noise, persons, physical constraints (e.g. amount of time, physical accessibility, comfort, cost) and surrounding information. **Social context** has information about "friends, neutrals, enemies, neighbours, co-workers, and relatives for instance". It also includes the roles played by the user, his status in these roles, the tasks he can perform in each role and the various sub-roles he can have. **Spatio-temporal context** describes aspects such as time, location, direction, speed, shape, track, place, clothes of the user, this is, the spatial extension of the environment and the things in it.

The two last dimensions are **economic techno-physical and societal contexts** and **historic context**. The first is more global than the environment context in the fourth dimension. It can include actual and global aspects like the H1N1 flu or the economic crisis. The sixth dimension involves all user's past actions.

3.2. Context uses

In this section are presented categories of uses of context in IR. From the authors of the reviewed context taxonomies, only Dey & Abowd (2000) has propose such an organization (which was later included in the Briconsouf & Newman's (2007) taxonomy). Their organization has three categories: presentation of information and services to a user, automatic execution of a service and tagging of context to information for later retrieval.

With the Dey & Abowd's (2000) categories in mind and with IR as this work's focus, the proposed top-level categories of uses of context in IR are: **indexing & searching**, **relevance feedback (RF) and query expansion**, **ranking**, **user interface**. These categories are components of an IR system were context may be used. The proximity of techniques used in the index construction and searching phases, stimulated their fusion in a single category. The proposed categories also map perfectly well to the categories defined by Dey & Abowd: the indexing & searching fits in the tagging category; the RF and query expansion may fit in the presentation of information and services (e.g. relevance feedback) or automatic execution of a service (e.g. implicit relevance feedback); the ranking fits in automatic execution of a service; and the user interface fits in the presentation of information and services.

The RF and query expansion category involves the processes of query refinement by the system, either fully automatically or with the help of the user. As defined by Manning et al. (2008), this category is divided in global and local methods. **Global methods** include query expansion/reformulation based on collection-independent knowledge structures (Efthimiadis 1996) like domain-specific thesaurus or general-purpose thesaurus (e.g.: WordNet), query expansion via automatic thesaurus generation and techniques like spelling correction. **Local methods**, like relevance feedback, pseudo relevance feedback and implicit relevance feedback, adjust the query with information from the documents that belong to the result set of the initial query. In relevance feedback the user marks returned documents as relevant or non-relevant and the system builds a better representation of the information need based on his feedback (Manning et al. 2008). Pseudo relevance feedback assumes the k ranked documents as relevant and implicit relevance feedback uses indirect sources of relevance.

The user interface category is also divided in two subcategories: the interface associated with the specification of the user's information need and the presentation of the result set. This last category is also divided in document surrogates (e.g. snippet - short summary of the document), query term hits within document (e.g. keyword-in-context snippets), categories for results set context and other type of strategies.

4. USE OF CONTEXT FEATURES IN IR

The proposed taxonomy was the basis for the analysis of a sample of contextual IR research papers. This sample is composed of 25 papers whose references are available at `http://www.carlalopes.com/papers_sample.pdf`. Papers' selection was made from a set of papers classified with the tag context (`http://www.citeulike.org/tag/context`) in *CiteULike*, a social web service for management of bibliographic references. In this list, papers related to IR, published in 2008, that made use of context features were included in our sample.

Each paper was examined towards the identification of: context definition adopted, context taxonomy exploited, context features used in the experience and their specific use. Only four papers introduced the adopted context definition and only one presented the underlying context taxonomy. Figure 3 has two pie charts where the left one shows the proportion of papers using each context feature's category and the right one shows the proportion of implemented context uses. In these graphs, each context feature (CF) and context use (CU) is represented by the numbers given in Figures 1 and 2.

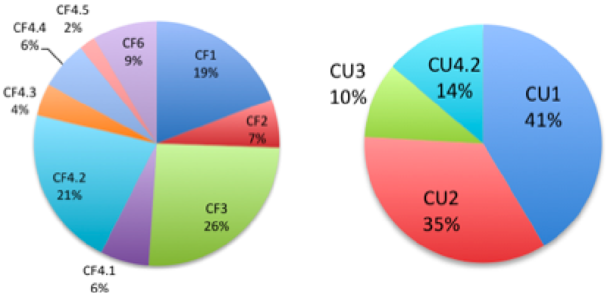


FIGURE 3: Context features and its uses in the set of analysed papers

The most used context features are the *interaction (session) context*, the *topic* and the *intra-object structures*. Interaction features range from desktop and web user behavior to users' tasks, actions and submitted queries. The context features in the topic category are as diverse as TREC topics' descriptions, context documents, domain thesaurus/ontologies and conceptual maps. *Indexing and searching* is the IR system's component where more papers employ context features, followed by the *RF and query expansion* with 35% of the papers.

5. FUTURE WORK AND ISSUES FOR DISCUSSION

This work is an initial phase of a PhD research that seeks to study how context features surrounding health information seeking and retrieval can affect the use of Health IR (HIR) systems and to apply these features in the improvement of these systems. The next step involves conducting an Health Information Seeking Behavior study to find the context attributes that matter for HIR applications. It will then be necessary to find ways to capture the identified context features and to define strategies to improve HIR involving the identified context elements.

Several issues are relevant for discussion that will, undoubtedly, be of great value to this PhD research. It would be interesting to discuss: ways to exploit shared contexts and contexts over time; evaluation methods and metrics of systems where users play a central role; common problems in IR experimental setups; ways to overcome these problems; testbeds suitable for the health area; envisioned research directions and pertinent research studies or literature to study.

6. ACKNOWLEDGEMENTS

I would like to acknowledge Cristina Ribeiro for reading and giving suggestions to this paper. This work is partially funded by Fundação para a Ciência e a Tecnologia under the grant SFRH/BD/40982/2007.

References

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval, *Canadian Journal of Information Science* (5): 133–143.

Bierig, R. & Göker, A. (2006). Time, location and interest: an empirical and user-centred study, *IIiX:*

Proceedings of the 1st international conference on Information interaction in context, ACM, New York, NY, USA, pp. 79–87.

Brézillon, P. (1999). Context in problem solving: a survey, *Knowl. Eng. Rev.* **14**(1): 47–80.

Briconsouf, N. & Newman, C. (2007). Context awareness in health care: A review, *International Journal of Medical Informatics* **76**(1): 2–12.

Cool, C. & Spink, A. (2002). Issues of context in information retrieval (ir): an introduction to the special issue, *Information Processing & Management* **38**(5): 605–611.

Dervin, B. (1997). Given a context by any other name: methodological tools for taming the unruly beast, *ISIC '96: Proceedings of an international conference on Information seeking in context*, Taylor Graham Publishing, London, UK, UK, pp. 13–38.

Dey, A. K. & Abowd, G. D. (2000). Towards a Better Understanding of Context and Context-Awareness, *CHI 2000 Workshop on the What, Who, Where, When, and How of Context-Awareness*.

Dourish, P. (2004). What we talk about when we talk about context, *Personal Ubiquitous Comput.* **8**(1): 19–30.

Efthimiadis, E. N. (1996). Query expansion, *Annual Review of Information Systems and Technology (ARIST)* **31**: 121–187.

Finkelstein, L. E. V., Gabrilovich, E., Matias, Y., Rivlin, E. H. U. D., Solan, Z. A. C. H., Wolfman, G. A. D. I. & Ruppin, E. (2002). Placing search in context: the concept revisited, *ACM Trans. Inf. Syst.* **20**(1): 116–131.

Göker, A. & Myrhaug, H. I. (2002). User context and personalisation, *ECCBR Workshop on Case Based Reasoning and Personalisation*.

Harper, D. J. & Kelly, D. (2006). Contextual relevance feedback, *IIiX: Proceedings of the 1st international conference on Information interaction in context*, ACM Press, New York, NY, USA, pp. 129–137.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory, *Journal of Documentation* **52**(1): 3–50.

Ingwersen, P. & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*, 1 edn, Springer.

Johnson, J. (2003). On contexts of information seeking, *Information Processing & Management* **39**(5): 735–760.

Kelly, D. (2006). Measuring online information seeking context, part 1: Background and method, *Journal of the American Society for Information Science and Technology* **57**(13): 1729–1739.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*, Cambridge University Press.

Mansourian, Y. (2008). Contextualization of web searching: a grounded theory approach, *The Electronic Library* **26**(2): 202–214.

Marchionini, G. (1997). *Information Seeking in Electronic Environments (Cambridge Series on Human-Computer Interaction)*, Cambridge University Press.

Saracevic, T. (1997). The stratified model of information retrieval interaction: Extension and applications, *Proceedings of the American Society for Information Science*, Vol. 34, pp. 313–327.

Saracevic, T. & Kantor (1988a). A study of information seeking and retrieving. ii. users, questions and effectiveness. *Journal of the American Society for Information Science* **3**(39): 177–196.

Saracevic, T. & Kantor (1988b). A study of information seeking and retrieving. iii. searchers, searches and overlap, *Journal of the American Society for Information Science* **3**(39): 197–216.

Saracevic, T., Kantor, P., Chamis, A. Y. & Trivison, D. (1988). A study of information seeking and retrieving. i. background and methodology, *Journal of the American Society for Information Science* **3**(39): 161–176.

Sato, K. (2001). Context sensitive interactive systems design: a framework for representations of contexts, *Hum. Comput. Interact.*, Vol. 2, pp. 229–241.

Winograd, T. (2001). Architectures for context, *Hum.-Comput. Interact.* **16**(2): 401–419.

Source Selection for Image Retrieval in Peer-to-Peer Networks

Daniel Blank
Media Informatics Group, University of Bamberg
Feldkirchenstr. 21, 96052 Bamberg, Germany
http://www.uni-bamberg.de/minf
daniel.blank@uni-bamberg.de

Abstract

With the emergence of web albums such as *Flickr.com* or *Picasa.com*, the amount of personal image collections administered on the web has increased dramatically. As a consequence, efficient storing, indexing and retrieval techniques are needed. Peer-to-peer (P2P) networks are an interesting solution to maintain large image collections. When performing a query on certain types of P2P networks, source selection is very important. In our scenario, compact summaries of each peer's image collection, which are known to other peers, are used to determine the most promising peers for a given query. These summaries have to address (1) date and time information, (2) textual information, (3) geolocations and (4) contend-based image features. The present paper outlines a large-scale image retrieval system, relying on data summaries and source selection strategies. While our scenario is based on a P2P system, we also describe how results can be transferred to other application domains such as distributed information retrieval (distributed IR) or tree-based index structures.

Keywords: Source Selection, Peer-to-Peer IR, Image Retrieval, Distributed IR, CBIR

1. INTRODUCTION

Web albums such as *Flickr.com* or *Picasa.com*, which offer storage capabilities for personal photo collections online, have become very popular in the last years. People upload their photos in order to share them with friends and to interact with each other e.g. by collaboratively tagging the photos.

Different criteria for image retrieval can be identified in such a scenario: mainly (1) date and time information, (2) tags and textual descriptions, (3) the geographic footprint, and (4) content-based image features describing e.g. colour or texture.

Our work addresses image search employing these criteria in peer-to-peer (P2P) networks. P2P scenarios for the administration of large image collections are attractive for multiple reasons. Photo collections can be stored locally on people's individual PCs. No expensive infrastructure has to be maintained by applying a scalable P2P protocol such as Rumorama [21] and remote computing power can be used to maintain the image collection. Users can decide which image features to publish in order for the corresponding images to become searchable without any need for crawling activities.

Our approach is based on—but not restricted to—Rumorama [21], a scalable P2P protocol building hierarchies of networks that are accessible by an efficient multicast. Its leaf networks behave like PlanetP networks [8]. In PlanetP, randomised rumour spreading assures that every peer knows summaries of all other peers' data in the network. The summaries provide the basis for source selection decisions, i.e. which peers to contact during query processing. While we examine peer summaries and source selection strategies for image retrieval in PlanetP-like middle-sized networks, we can easily extend this to large-scale Rumorama-like P2P networks.

This paper is organised as follows. Section 2 gives a brief overview on related work. In section 3, the more general applicability of peer summaries in different application scenarios is discussed.

Section 4 describes our main approach, different summary-types and source selection stategies. In section 5, we conclude with an outline of challenging research questions for future work in order to design a large-scale, distributed image retrieval system based on source selection.

2. RELATED WORK

To our best knowledge, there are no multi-feature P2P information retrieval (IR) systems that allow for text-based and content-based image retrieval (CBIR) additionally employing temporal and geographic metadata. In our scenario, in order to support image retrieval based on these criteria, a peer will have to maintain and distribute at least four different summary types. For the summarisation of linear time and date information, we assume that our summaries presented in section 4 are directly applicable. Traditional histogram techniques (cf. [17]) as well as techniques used for aggregating sensor data (cf. e.g. [24]) can also be applied. Summarising the textual annotations and descriptions for image retrieval will be part of future work. Within this paper in section 4, we focus on the summarisation of multi- and high-dimensional data.

In general, P2P IR systems can be classified into several groups (cf. [6])[1]. Systems of the first group follow a semantic query routing approach based on peer summaries. Routing Indices [9] are among the first approaches presented in literature belonging to this group. Based on summary information of neighboring peers that is aggregated along multiple hops, a peer routes queries towards the direction of peers potentially containing relevant documents w.r.t. the query. In order to restrict the size of peer summaries, topics are indexed rather than individual terms.

As opposed to Routing Indices, which follow a multi-hop semantic routing approach, PlanetP [8] and its scalable extension Rumorama [21] apply single-hop semantic routing. Therefore, summaries are sent to all peers in a (sub-)network. Summaries and source selection strategies for text data based on Bloom filters are analysed in [8, 11].

The single-hop semantic query routing approach originally comes from distributed IR (for references see e.g. [22, 25]). Many of the approaches proposed in literature assume that it is feasible to base routing decisions on term frequency or term weighting information which is available for all terms of a particular resource. But, summaries within P2P IR systems need to be more space efficient than the ones designed for distributed IR, because of limited bandwith capacities and the frequent joining and leaving of peers, often with updated document collections. Therefore, most of the traditional distributed IR approaches are not directly applicable.

The second group of P2P IR systems are semantic overlay networks (e.g. [18]) where the content of a peer's data defines its place within the network topology. Peers are organised by semantic clusters and within query execution the query is routed to the most promising cluster(s). Here, the indexing of multiple feature types, e.g. textual information as well as image content, would require the definition of a similarity between peers combining textual and image content information. Alternatively, several overlays might be maintained inducing a higher maintenance effort.

A third class of P2P IR systems is represented by distributed indexing structures with distributed hashtables (DHTs) as its most prominent class member. Minerva [2] has been designed for the administration of text documents, where term statistics are indexed in a DHT. Every peer is responsible for a certain set of terms. Novak et al. have presented a large-scale CBIR architecture [23] based on a DHT. Within DHTs, indexing data of a peer's content is transferred to remote peers with every peer being responsible for a certain range of the feature domain of an individual feature. Presumably, for example, correlations between geographic information and image content are difficult to exploit. If we e.g. assume an image from the Sahara Desert with shades of beige sand and blue sky, different peers might be responsible for indexing the geographic and the image content information. Therefore, when distributing the indexing data of the Sahara image, querying for it, or removing it from the network, two different peers (at least)

[1] In the following, we only describe *true* P2P IR systems. Super-peer scenarios, where some peers are chosen to perform certain tasks because of increased capabilities of their resources, are not analysed in more detail. Nevertheless some of the strategies presented in section 4 can also be applied in a super-peer scenario (for a brief description see section 3).

have to be contacted. Even with only one feature type being indexed (e.g. text in the case of Minerva), the frequent joining and leaving of peers leads to an increase in network traffic as term statistics are transferred to or removed from remote peers.

In order to summarise collections of personal photos for distributed CBIR, other approaches than the ones presented in section 4 are also possible. Earlier work in distributed CBIR follows a clustering approach. Chang et al. [7] create summaries of remote databases based on image templates, i.e. feature vectors of reference images, sampled from remote databases. Hierarchical clustering is applied in order to obtain a set of cluster centroids. The authors present two approaches to source selection, the first based on statistical information and the second based on histograms. The latter is similar to our approach, being different for example w.r.t. the sampling phase used for obtaining the centroids, the clustering process itself, the way how histograms are computed, and especially w.r.t. the number of centroids used for computing the summaries.

Berretti et al. [3] apply a special form of hierarchical clustering to the image features of a remote database. With the use of a threshold similarity it is possible to adjust the number of centroids, i.e. the granularity and size of the resource descriptions. The resource descriptions consist of the centroids themselves. In [14] we also applied a local clustering technique using mixtures of Gaussians. For two-dimensional geographic data, cluster hulls have been proposed in order to summarise sets of geographic coordinates of an individual resource by several convex hulls [16].

In general, we expect approaches explicitly transferring centroids to be less space efficient than our approach presented in section 4, especially for high-dimensional feature vectors. Histogram-techniques implicitly using centroid information seem to be more promising.

3. APPLICABILITY OF SUMMARIES AND SOURCE SELECTION STRATEGIES

We analyse summaries for PlanetP-like P2P systems (cf. section 4). Nevertheless, these summaries are not restricted to PlanetP and related protocols that make use of single-hop semantic routing. Within multi-hop semantic routing networks, summaries will have to be aggregated along multiple hops. Summarisation of peer content is also needed in semantic overlay networks in order to derive a peer's place within the network topology. In this case, it is necessary to define a similarity between peer summaries so that peers with similar summaries can be grouped together into “clusters of interest”.

Many P2P IR protocols rely on super-peers. Typically, they are characterised by increased storage or bandwith capabilities w.r.t. “normal” peers. Super-peers also tend to stay in the system for most of the time. Query routing is usually performed amongst super-peers and “normal” peers transfer their indexing data to responsible super-peers. Only if a document is transferred from a “normal” peer to another, this is done at the peer level without involving super-peers. Therefore, our summaries can also be used in a super-peer scenario for content summarisation of “normal” peers and as the basis for query routing amongst super-peers.

We believe that our work can also be beneficial for the design of tree-based index structures. Summaries in the P2P context with their enforced space limitations are similar to approximations maintained in inner nodes of a tree. Indexing structures e.g. relying on minimum bounding rectangles such as the R-tree [15] and its variants can provide the basis for summary construction of geographic data [5]. Signatures / Bloom filters are used to summarise textual data (cf. e.g. [8, 10]).

During the last years, local, content-based image features like e.g. SIFT [19] have become popular. In the case of SIFT, an individual image is characterised by several hundred *128*-dimensional feature vectors. This poses increased challenges on indexing techniques. We believe that our summarisation and source selection techniques presented in section 4 can be adapted to summarise and index SIFT features, both on a per-image as well as on a per-collection basis.

After more than a decade of P2P research, application scenarios, where P2P IR technology is successfully used, are still missing. At the moment, cloud computing—with some respect oppositional to the P2P concept—seems to be appealing. Large computer infrastructures with free capacities are used in order to provide software as a service as well as storage capacities online. Following this trend, it might become reasonable for companies and individuals to use several different storage services in parallel because of pricing and availability reasons. Therefore, source selection based on compact data summaries might become important.

Summarisation of documents as well as source selection strategies are of course also important in traditional client/server applications. If we think of market situations with many buyers and sellers, source selection strategies based on summarisations of indexing data might provide a benefit for all participants. There is for example a huge amount of photo agencies and services providing images that are subjective to licence conditions and charge. For clients searching via traditional web search, it is difficult to identify the most promising service providers in order to browse through their sites. Also the design of a meta search engine might be difficult since textual metadata is often stored in content management systems, largely hidden for traditional web crawlers. The extraction of content-based image features is hampered as images are often only available in a very small resolution, with many of them being modified by watermarks. In such a scenario, providing indexing data for a centralised indexing broker might be beneficial for service providers to gain attention. Additionally, users might benefit from this service, as automatic source selection would prevent them from browsing too many irrelevant sites. Brokers might gain revenue—similar to traditional search engines—through advertisements and/or online auctions with service providers bidding for high ranks.

Personal metasearch is a novel application domain of distributed IR [25]. All of a user's online resources are summarised and metasearch is provided based on summaries integrating heterogeneous resources that largely vary in size. This is oppositional to many traditional distributed IR scenarios. It is therefore important to make use of source selection strategies which—in the case of varying resource sizes—do not prefer to contact large collections as this is not always a good choice. Additionally, personalised online resources consist of different types of data with high update frequencies (email accounts, Web sites, photo and video sharing communities, local databases etc.). We therefore believe that our summaries designed for P2P networks can also be applied for personal metasearch, as we require them to be selective (also being able to identify promising small peers administering few documents) and space efficient (because of the dynamic nature of P2P systems). Additionally, we target on multiple summary types for temporal, geographic, textual and content-based (meta)data useful for the summarisation of textual, image, audio and video content.

4. SOURCE SELECTION STRATEGIES FOR IMAGE RETRIEVAL

We have analysed cluster histograms for summarising collections of real-valued feature vectors [12]. In order to compute cluster histograms, all peers need to know a unique set of cluster centroids in feature space. A peer joining the network will try to obtain the centroids from peers already present. If peer p has obtained the set of centroids $C = \{c_i | 1 \leq i \leq \kappa\}$, peer p assigns every feature vector of its local collection to the closest centroid c_i, according to a given distance measure. The same distance measure is used by all peers in the system. So, peer p computes a histogram that assigns to every centroid c_i the number of feature vectors closest to c_i. The cluster histogram summarises the data collection of peer p. Peers publish their summaries by randomised rumour spreading.

Obviously, the combination of summaries and source selection strategies is crucial for the performance of query processing. In [12], three source selection strategies have been evaluated. The most promising strategy sorts the cluster centroids c_i in a list *L* in ascending order according to their distance to the query. The first element out of *L* corresponds to the centroid of the cluster that is closest to the query. Peers with many documents inside this so called query cluster are ranked higher than peers with less documents in the query cluster. If two peers share the same amount of documents in the currently analysed cluster, the next element out of *L* is chosen and the two peers are recursively ranked w.r.t. the number of documents within the current cluster.

As an example, let us assume three centroids ($\kappa = 3$), two peers p_a and p_b with corresponding summaries $s_a := (5, 8, 3)$ and $s_b := (7, 8, 2)$. We assume that centroid c_2 is the closest and c_1 is the second closest centroid w.r.t. query feature vector q. Both peers have assigned 8 image feature vectors to the cluster represented by centroid c_2. But, out of peer p_b's image feature vectors 7 are closest to c_1, compared to 5 feature vectors that are closest to c_1 out of peer p_a's image collection. As $7 > 5$, peer p_b is ranked higher than peer p_a and contacted before p_a during query processing.

An important finding in [12] is that a distributed clustering for computing the set of centroids might be dispensable. A random selection of feature vectors out of the data collection that is administered in the P2P system can be used as centroids. Their usage results in a minimal decrease in retrieval performance at the same time making distributed clustering obsolete. In internal experiments also other variants of distributed clustering like fuzzy *k*-Means clustering and self-organising maps could not improve retrieval performance.

In [13] we analysed the performance of cluster histograms for a large number of distance measures and image features. It became clear that our source selection strategy is influenced by the curse of dimensionality.

The creation of cluster histograms presented so far needs some global knowledge. The peers present in the system must agree on the set of centroids. This does not affect the usability of cluster histograms, but it restricts the adaptivity of our approach, as an update of the centroids becomes expensive. Approaches where peers do not have to agree on the set of centroids are therefore desirable. Local clustering techniques and Gaussian mixture models (GMMs), that do not rely on global knowledge when computing the summaries, are analysed in [14]. When performing local clustering, a peer computes a small number of local clusters and their centroids are published as summaries. GMMs model the point density distribution as superposition of Gaussians with different means and covariances. GMMs outperform local clustering. But, in our experiments relatively small cluster histograms (256 bins or clusters) with globally distributed cluster centroids are superior to GMMs in terms of retrieval performance. Therefore we optimised the former strategy as explained in the following.

Our new summaries—called highly fine summaries (HFS)—evolve from the cluster histograms described earlier by varying κ, the number of cluster centroids [4]. We increase the number of centroids from 256 to e.g. 16,384 or even more. This offers several benefits for our scenario:

- Retrieval performance is improved since the data space is partitioned in a more fine-grained way.
- At the same time, the costs for distributing the summaries only increase moderately as we compress the summaries. We use runlength encoding which allows us to substantially reduce summary sizes. This is possible since with large numbers of centroids, many of the small peers (i.e. peers administering few images) will compute summaries with some histogram entries set to very small values (often 1), but most of the histogram values will stay 0 as no image is assigned to the corresponding centroid.
- Administration overhead for distributing the centroids even decreases as we distribute the set of centroids with software updates. Within our experiments we showed that if we choose the centroids from a different, disjoint collection within the same application domain (we use images from *Flickr.com*), average retrieval performance is not affected.

HFS are designed for high-dimensional image feature vectors. In [5] we analyse the applicability of modified HFS—now called ultra fine summaries (UFS)—in the context of summarising sets of geographic coordinates. Instead of cluster histograms, we use bit vectors, where bit i is set, if for any image out of a peer's collection, centroid c_i is the closest. Otherwise, the corresponding bit remains zero. We have evaluated the usage of UFS for point queries against summarising a peer's geographic footprint by either a minimum bounding rectangle or a grid-based, binary index. UFS show the best performance in terms of selectivity. Therefore we will evaluate source selection strategies for *k*-nearest-neighbour queries based on UFS in future work.

5. CHALLENGES AND FUTURE DIRECTIONS

HFS/UFS seem promising for summarising content-based image features as well as sets of geographic coordinates. Other important criteria for image retrieval are time and date information and textual information. We believe that HFS/UFS might also be promising for summarising text data. Therefore, we will compare HFS/UFS with traditional Bloom filter approaches. Spectral Bloom filters might be employed to encode term frequency information [11] or impact information [1] into the summaries. In order to do so, a large, distributed test collection is needed that offers a realistic distribution of text documents to peers. As our main application scenario is image retrieval, we might use image collections from Flickr.com, being crawled together with textual descriptions and tags.

When compressing HFS (i.e. histograms of integer values) of big peers, the summaries might become large. In the future, we will analyse the differences between UFS and HFS in more detail. Retrieval performance might decrease as UFS do no longer maintain frequency information about how many of a peer's images are closest to a certain centroid. At the same time, UFS allow us to use more centroids as only a single bit is needed to encode, if any of a peer's images is closest to a certain centroid. Additionally, UFS might be beneficial for compression as bit vectors with many bits set to 1 are still suited for compression.

In order to restrict the summary size of big peers, it is necessary that a peer is granted a maximum amount of space to encode its summary. The size of a summary in the case of HFS might therefore be chosen depending on *n*, i.e. the number of documents that a peer administers, e.g. by multiplying a basic summary size with *1+log(n)* or similar factors. Another approach for restricting the size of the summaries might be to hierarchically partition the data space. A peer can then choose the number of centroids according to this space partitioning so that the overall summary size does not exceed a given upper bound.

We have to derive a better stopping criterion determining if it is promising to contact further peers or to stop query processing. Currently we stop after having contacted a certain fraction of peers. This value is determined through experiments. In future, we will analyse and adapt solutions originally proposed for textual data [8, 20] and design new ones if necessary.

Within our approach, we use a secondary collection from which the centroids are chosen. Currently we choose them randomly. But, applying clustering techniques or specialised selection techniques might be beneficial in order to gain in retrieval performance. A distance matrix *D* with pairwise distances between centroids may provide additional benefits. The ordering of indexes within HFS/UFS depending on *D* or based on some type of space filling curve might be beneficial for compression. Furthermore, using *D* in addition with the triangular inequality might be helpful in order to derive algorithms for a centralised indexing structure based on our HFS/UFS approach.

The secondary collection might also be used in order to apply dimensionality reduction. In [13], we have seen that the quality of the source selection decreases with increasing dimensionality of the underlying feature space. Therefore, the secondary collection and the centroids might provide a basis for local dimensionality reduction. We will therefore compare the effects of distributed PCA (PCA: principal component analysis) with local PCA based on the secondary collection as well as other local techniques for dimensionality reduction.

As discussed earlier in section 3, using modified versions of HFS/UFS in combination with adapted source selection strategies seems also promising in order to index local feature descriptors with several hundred feature vectors per image – both on a per-image as well as on a per-collection basis.

REFERENCES

[1] Anh V. N. and Moffat A. (2005) *Simplified similarity scoring using term ranks.* Proc. of 28th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 226-233, Salvador, Brazil.

[2] Bender M., Michel S., Weikum G. and Zimmer C. (2005) *The MINERVA Project: Database Selection in the Context of P2P Search.* GI-Conf. on Database Systems in Business, Technology and Web, pp. 125-144, Karlsruhe, Germany.

[3] Berretti S., Del Bimbo A. and Pala P. (2004) *Merging Results for Distributed Content Based Image Retrieval.* Multimedia Tools and Applications, 24(3), pp. 215-232, Kluwer.

[4] Blank D., El Allali S., Müller W. and Henrich A. (2007) *Sample-based Creation of Peer Summaries for Efficient Similarity Search in Scalable Peer-to-Peer Networks.* 9th ACM SIGMM Int. Workshop on Multimedia Information Retrieval. pp. 143-152, Augsburg, Germany.

[5] Blank D. and Henrich A. (2009) *Summarizing Georeferenced Photo Collections for Image Retrieval in P2P Networks.* Int. Workshop on Geographic Information on the Internet, Toulouse, France.

[6] Blank D., Müller W. and Henrich A. (2008) *Designing Benchmarks for the Evaluation of Peer-to-Peer Information Retrieval Systems.* In: Baumeister J., Atzmüller M. (Edts.), GI Workshop: Lernen, Wissen & Adaptivität, LWA 2008, Würzburg, Germany.

[7] Chang W., Sheikholeslami G., Wang J. and Zhang A. (1998) *Data Resource Selection in Distributed Visual Information Systems.* IEEE Trans. on Knowl. and Data Eng., 10(6):926–946.

[8] Cuenca-Acuna F. M., Peery C., Martin R. P. and Nguyen T. D. (2003) *PlanetP: Using Gossiping to Build Content Addressable Peer-to-Peer Information Sharing Communities.* Proc of the 12th IEEE Int. Symp. on High Performance Distributed Computing. pp. 236-246, Seattle, WA, USA.

[9] Crespo A. and Garcia-Molina H. (2002) *Routing Indices For Peer-to-Peer Systems.* Proc. of 22nd IEEE Int. Conf. on Distributed Computing Systems, pp. 23-32, Vienna, Austria.

[10] Deppisch U. (1986) *S-tree: a dynamic balanced signature index for office retrieval.* Proc. of 9th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 77-87, Pisa, Italy.

[11] Eisenhardt M., Müller W. and Henrich A. (2004) *Spektrale Bloom-Filter für Peer-to-Peer Information Retrieval.* Lecture Notes in Informatics, No. P-51, pp. 44-48, Ulm, Germany – ISBN 3-88579-380-6.

[12] Eisenhardt M., Müller W., Henrich A., El Allali S. and Blank D. (2006) *Clustering-based source-selection in summary-based P2P networks.* Proc. of 8th IEEE Int. Symp. on Multimedia, pp. 823-830, San Diego, CA, USA.

[13] El Allali S., Blank D., Eisenhardt M., Henrich A. and Müller W. (2007) *Untersuchung des Einflusses verschiedener Bild-Features und Distanzmaße im inhaltsbasierten P2P Information Retrieval.* GI-Conf. on Database Systems in Business, Technology and Web, pp. 382-396, Aachen, Germany.

[14] El Allali S., Blank D., Müller W. and Henrich A. (2008) *Image Data Source Selection Using Gaussian Mixture Models.* Proc. of 5th Int. Workshop on Adaptive Multimedia Retrieval, pp. 170-181, Paris, France, Springer.

[15] Guttman A. (1984) *R-Trees: A Dynamic Index Structure for Spatial Searching.* Proc. of ACM SIGMOD, pp. 47-57, Boston, MA, USA.

[16] Hershberger J., Shrivastava N. and Suri S. (2008) *Summarizing spatial data streams using ClusterHulls.* ACM Journal of Experimental Algorithmics, 13 (2.4-2.28).

[17] Ioannidis Y. (2003) *The History of Histograms (abridged).* Proc. of 29th Int. Conf. on Very Large Data Bases, pp. 19-30, Berlin, Germany, ACM.

[18] King I., Ng C. H. and Sia K. C. (2004) *Distributed content-based visual information retrieval system on peer-to-peer networks.* ACM Trans. Inf. Syst. 22(3). pp 477-501.

[19] Lowe D. G. (2004) *Distinctive image features from scale-invariant keypoints.* Int. Journal of Computer Vision, vol. 60, pp. 91-110.

[20] Lu J. and Callan J. (2004) *Federated search of text-based digital libraries in hierarchical peer-to-peer networks.* Proc. of SIGIR Workshop on Peer-to-Peer Information Retrieval, Sheffield, UK.

[21] Müller W., Eisenhardt M. and Henrich A. (2005) *Scalable summary-based Retrieval in Peer-to-Peer Networks.* Proc. of the ACM Int. Conf. on Information and Knowledge Management, pp. 586-593, Bremen, Germany.

[22] Nottelmann H. and Fuhr N. (2003), *Evaluating different methods of estimating retrieval quality for resource selection.* Proc. of 26th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 290-297, Toronto, Canada.

[23] Novak D., Batko M., Zezula P. (2008) *Web-scale system for image similarity search: When the dreams are coming true.* Workshop on Content-Based Multimedia Indexing, pp. 446-453, London, UK, IEEE.

[24] Shrivastava N., Buragohain C., Agrawal, D. and Suri, S. (2004) *Medians and beyond: new aggregation techniques for sensor networks.* Proc. of 2nd ACM Int. Conf. on Embedded networked sensor systems. pp. 239-249, Baltimore, MD, USA.

[25] Thomas P. and Hawking D. (2009) *Server selection methods in personal metasearch: a comparative empirical study.* Information Retrieval, Springer, *http://www.springerlink.com/content/y3704552037635jj/*

Intelligent media indexing and television recommender systems

Tamas Jambor
Department of Computer Science
University College London
t.jambor@cs.ucl.ac.uk

This paper presents a state-of-art review on recommender systems that has been identified as possible areas of my research. It describes the current generation of collaborative filtering methods which are usually classified into three main categories: item-based, user-based and hybrid methods. The paper considers these methods to be applied to digital television providing recommendation for viewers. Personalised television is predicted to be the next step in the evolution of television which might reshape the whole landscape of mass media. The paper also identifies anticipated problems in the domain of recommender systems which includes indexing, collaborative filtering, ranking problems and possible research directions to solve these problems. Finally, these challenges are considered in the domain of personalised television which has its own inherent shortcomings.

Collaborative filtering, recommender system, multimedia retrieval, personalised television

1. MOTIVATION FOR RESEARCH

The digital switchover will be completed by 2012 in the UK it implies that all households will have access to some kind of digital television content. Digital television offers a wider range of channels where the user can access a large amount of content thought terrestrial, satellite or cable broadcast. Video on demand (VOD) and electronic program guide (EPG) services are part of this expansion, which help to make digital television more flexible and provide easy to access information.

In economy Anderson (2008) introduced the concept of long tail distribution, it shows that retailers sell relatively large quantities from small number of popular items and sell large number of items which are not that popular in smaller quantities. This distribution can be applied to multimedia item, since multimedia items can also be considered as products. This suggests that there is a direct correlation between users' interest and the popularity of multimedia items (FIGURE 1), that means popular items meet only a small part of the populations' taste. The reason why popular items get popular despite that they are only meet the interest of a small proportion of the population can be explained by the nature of mass media, people tend to watch programs that are available in peak time and usually heavily advertised, which they would not normally watch. Therefore, mass media pulls together people who would not normally find themselves together (Anderson 2008).

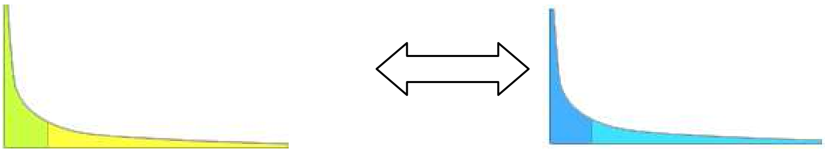


FIGURE 1: User's interest and the popularity of content (illustration)

However, VOD services enable non linear access to content, so it is possible to serve viewers who have less popular taste, which cannot be satisfied by conventional television. This would reduce the popularity of currently popular items and create a flatter distribution. However, the variety of content would inevitable create information load so users might find it difficult to find programs that they wish to watch, and sometime they spend more time browsing channels or browsing the digital library than watching actual programs. In addition to that users might not know what exactly they are interested in, therefore it is not possible to them to search explicitly for content. To overcome this personalisation systems aim to close the gap between potential interests and available items by suggesting items that match to that interest. However, providing reliable recommendation gets more problematic as the system deals with the 'long tail' of the data.

Personalisation is a powerful technique which might reshape the whole landscape of mass media, personalised television is predicted to be the next level in the evolution of interactive television. Personal TV suggestion engines (Wang et al. 2008) have appeared in recent years, offering personalised TV listings, personalised content channels (e.g. beeTV), etc. However, these services only adopted a simplified version of online recommendation system, because of the inherent problems with digital television. They usually offer personalised services based on genre or program types (e.g. thriller, comedy) manually set by the user, only very few of them takes into account previous user preferences.

Since the technology enables service providers to keep a log of user activity, this information can be used as a basis for research to develop a retrieval tool which would provide personalised recommendation. In order to provide efficient recommendation this project also aims to develop an indexing tool that will create sufficient data for retrieval.

The proposed system would contain five main modules (FIGURE 2). A user profile module would be in charge of monitoring and logging users' behaviour and storing this data in the database. It would also modify user data based on pieces of information on how users perceive items that are presented to them (profiling). The other information gathering module would create metadata for multimedia content. This might include a collaborative tagging system and an automatic metadata retrieval tool.

The main part of the system is the recommendation engine. This module would return a list of recommended items for any given user using the two datasets (users, items). It also receives feedback from users that will directly affect the procedure. This list will be passed to the post-filtering module which would filter recommended items based on contextual information. The last part of the process would rank the results by popularity, date, etc. and present it to the user.

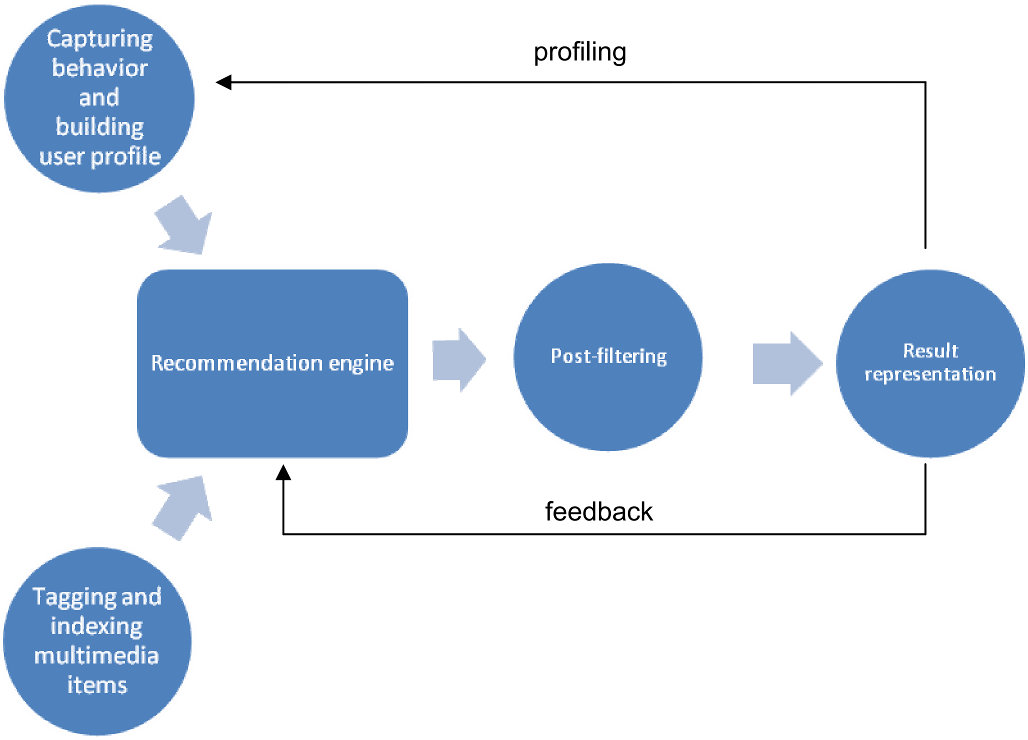

Capturing behavior and building user profile
profiling
Recommendation engine
Post-filtering
Result representation
feedback
Tagging and indexing multimedia items

FIGURE 2: Social information retrieval that combines user intelligence with the content features

The rest of the paper will present a brief overview of the work that is relevant to this project in indexing and multimedia retrieval. The second part will focus on the anticipated research problems in general and in the television recommendation system domain.

2. BACKGROUND AND RELATED WORK

Recommender systems are a means of personalisation providing users with personalised recommendation that would possibly suit users needs. Recommender systems are used in a broad range of applications and web services.

2.1 Collaborative filtering

One of the most popular techniques that are used for recommender systems is collaborative filtering (Herlocker et al. 1999). It enables the system to build coherent communities based on shared taste and behaviour. These methods use the opinion of users to help individuals to identify content of interest from a large set of choices, which otherwise would be problematic to find. The basic idea is that recommendations are provided on the basis of user profiles and on the metadata of particular items. The result is a set of recommended items that have not been known to the user before, but match to his or her taste. These techniques are based on two basic assumptions, first of all, users who like the same documents are assumed to have the same interest. Secondly, it is assumed that users taste is relatively constant.

User-based filtering techniques (Wang et al. 2006) concentrate on finding similar users to the active user. First a set of nearest neighbours of the target user are computed. This is performed by computing correlations or similarities between user records and the active user. Then, different methods are used to combine the neighbours' item ratings to produce a prediction value for the target user on unrated items. The major problem with this approach is the bottleneck problem, the complexity of the system increases as the number of users grows which could reach an unmanageable number of connections to compute in large commercial systems.

Item-based collaborative filtering (Sarwar et al. 2001) aims to find items which are similar to a particular user's preferences. This algorithm attempts to find similar items that are co-rated (or visited) by different users similarly. This is done by performing similarity calculation between items. Thus, item-based algorithms avoid the bottleneck in user-user computations by first considering the relationships among items.

It has been shown (Adomavicius and Tuzhilin 2005) that hybrid recommender systems are more efficient than systems based purely on item-based or user-based collaborative filtering. In addition to that data sparsity is considered one of the main reasons why these systems perform poorly. Wang et al. (2006) proposed a hybrid system where user-based and item-based collaborative filtering approaches are unified using probabilistic fusion framework, it is showed that this system performs better even with sparse data.

2.2 Indexing

In order to build a useful database for retrieval, pieces of information about multimedia items are retrieved from the EPG provider. In some systems (Tsunoda and Hoshino 2008) automatic metadata expansion is used to convert this information to metadata. However, systems which are aimed to unite content from different providers might find it problematic to obtain the same set of information for each multimedia item, since different providers have different internal data structures.

Alternatively, collaborative tagging processes offer techniques to add some extra information in order to enrich the metadata of a multimedia item. Concerns were raised on the reliability of this method, since users might assign tags in an uncontrollable manner, which would result unsystematic metadata. However, some researchers (Golder and Huberman 2005) found stable patterns in the structure of these tags. It is suggested that the proposition of tags becomes fixed if enough users assigned tags to a particular item. After a threshold is reached, 'each tag's frequency is a nearly fixed proportion of the total frequency of all tags used' (Golder and Huberman 2005, p. 6).

3. RESEARCH CHALLENGES

3.1 Lack of semantic descriptions

Since the traditional way of multimedia retrieval can be problematic, because of the gap between low-level features of the multimedia items and the semantic symbols that is used for retrieval. Recently researchers begin to explore social multimedia signals (Chang 2008) that can be fused with traditional features of content-based multimedia retrieval which might result more effective multimedia retrieval. Interaction of users can be used to identify low level features, for example adjectives that are used in a descriptive manner might refer to some of the features of the multimedia item.

3.2 User interests' predictions

Collaborative filtering is a very well research area, because it can be applied to a wide range of problems, for example predicting user's taste in retail, music, cinema etc. It can also offer prediction to estimate if a particular product would be successful or not. Therefore providing efficient recommendation or prediction is a key to success for commercial applications.

One of the main problems that might arise in online recommendation systems is the so-called cold start problem (Schein et al. 2002). Until a certain point, there is not enough data to make sufficient recommendation, in order to overcome this, systems usually set a limit (e.g. Netflix) which has to be reached in order to provide enough data for the recommendation engine. On the item level this problem appears differently. Since old items have more rating than new ones, recommender systems tend to be biased towards the old and have difficulty showing the new. This might be the problem, since users tend to prefer new items over old ones. One of suggestions to solve this problem (Lee et al. 2009) is to take temporal information into account, temporal information includes the launch time of the item and the time difference between the item was rated and the launch time.

One of the assumptions discussed above was that users' taste does not change significantly over time (Boyer and Brun 2007), however, some recommender systems consider users' taste within a shorter time frame, and assume that in a longer term users' profiles can expire (Tsunoda and Hoshino 2008). In addition to that a system can separate a core taste of the user and a temporary taste which can be computed by comparing similar items in users' profiles. Then the core taste could be used in a longer period of time while the temporary taste can expire if the user becomes interested in items that do not match her previous temporary taste.

Popular movies and unpopular movies has a shape of distribution shown in FIGURE 3 which make the movie easy to predict, because the majority of users grouped together either on the positive side or the negative side of the scale.

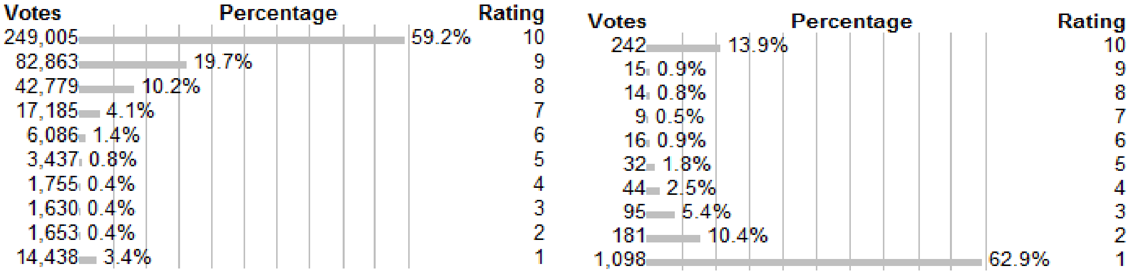


FIGURE 3: Popular (The Shawshank Redemption), unpopular (The Skydrivers) distribution (source: IMDB.com)

In FIGURE 4 both movies have the same average rating (5/10), however, the distribution is very different. Items which people either love or hate are more problematic to predict. This can be accounted to fact that these movies have only good and bad ratings and nothing in between that makes it problematic to match it to users' profiles. For example users who loved or hated *Michael Moore hates America* clearly based their rating on political orientation rather than previous taste so from the point of known recommender algorithms users behave unpredictably. Also, unpredictable items might not linked very closely to the rest of the system, because they represent a new style or represent a theme that is unique, which also makes it difficult to predict whether it will be liked or not.

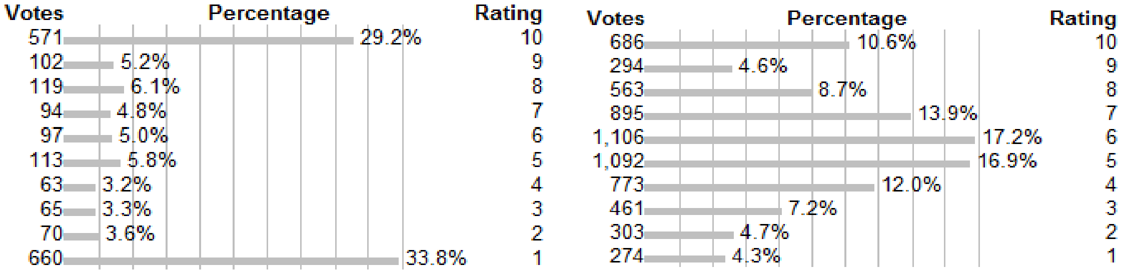


FIGURE 4: Love or hate distribution (Michael Moore hates America), neutral distribution (Captain Ron) (Source: IMDB.com)

Current recommender system do not differentiate between item and item in their internal data structure, thus these differences cannot be taken into account during the process of recommendation. Therefore, we aim to investigate this matter further by indentifying items that should be considered differently based on the distribution of the votes or other criteria (e.g. other systems found them unpredictable) and developing strategies that can be tailored to threat a particular item individually and use strategies that fit best for that item to provide efficient recommendation.

Since previous votes tend to influence future votes and users' taste changes over time, it is possible that the shape of the distribution changes as the item receives more votes. FIGURE 5 illustrates the idea, each column represents the vote distribution after a particular number of votes. It shows that the semantic orientation of the vote distribution can change significantly over time as movies receive a high volume of votes that differ from the current distribution. This measure would represent the fluctuation in users' taste over time. In this case (FIGURE 5) users tend give rating four and five more to the movie *Silkwood* compared to the initial distribution (after 500 votes). That suggests that users like the movie more than they initially liked it. In this case it is a shift towards higher votes, but there could also be a two directional shift, that is a shift from the middle towards both ends of the scale.

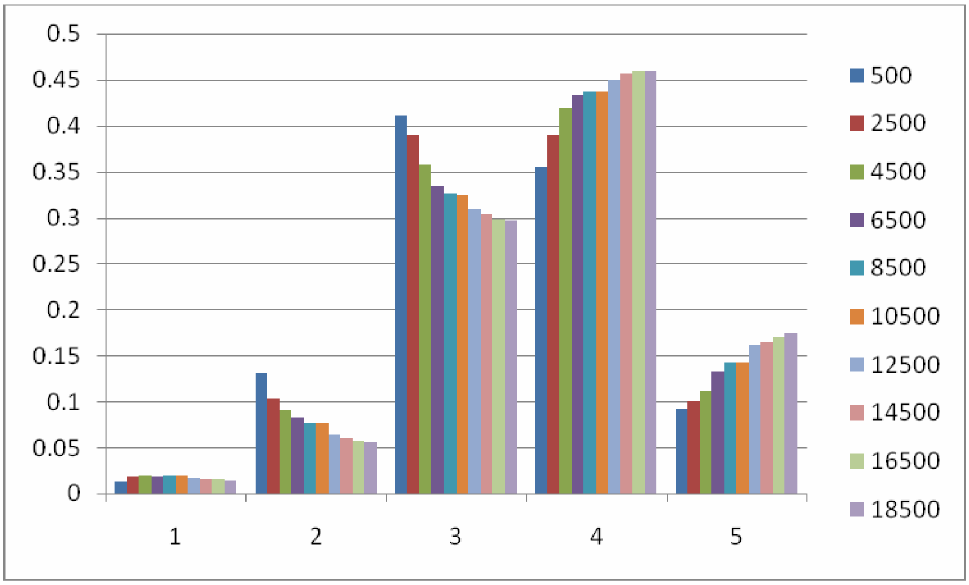


FIGURE 5: Changes of distribution of *Silkwood* (Source: Netflix dataset)

3.3 Contextual information and ranking problems

A recommender engine would return many items that are potentially interesting to the user, for that reason it is suggested to apply a post-filtering algorithm which would return only a few items that can be displayed on screen. Also, recommender systems consider user profiles as a homogeneous system and take into account items equally. However, in reality taste is more a multidimensional system which might change with time, influenced by other users etc. For example a user might prefer to look different items at the weekend than at weekdays or his or her taste might change slightly after they watched a particular movie. These subtle modifications of the system could be achieved by post-filtering results returned by the recommender engine. Contextual information (e.g. user’s age, favourite genre) (Adomavicius et al. 2005) can be used to filter results returned by the recommendation engine. Since these parameters can change quicker than the general taste of the user, contextual information should be considered separately, not as part of the recommender system. These techniques are useful to differentiate the core taste and the temporary taste of the user to threat time as a separate dimension.

General sentiment score is an indication of the general popularity of a multimedia item that can be obtained from different sources using NLP (Natural Language Processing) techniques. For example the frequency of particular adjectives that occur near to a multimedia item (e.g. in comments, in reviews) can give an indication of how well that item is perceived by the general public. This piece of information can be used to rank the search results and give the user the most popular items first within the domain of their preference. This information would provide a different way of measuring data compared to ratings, since both measurements express sentiment towards an item in different ways. However, it is very likely that these measurements do not differ significantly, since they only represent different ways of expressing opinion.

3.4 Program clustering on digital television

In addition to problems anticipated in online recommender system, recommender systems of interactive television suffer from several inherent problems. It includes primitive user interface, poorer access to information, the services providers’ monopoly on supplying software for the digital box (as opposed to a more comparative environment on the web).

Since a television set might be shared between people who are living in a same household, it is problematic to determine who is watching a program at any given time. One solution might be to set up a user profile system, which would enable users to log into their TV sets. From that point the system would function as a standard recommendation system. The other solution would be to take contextual information into consideration, for example timestamp could be used to determine who is watching the television at any given time based on the assumption that similar type of programs are likely to be watched on the same television set at the same time of each day (e.g. news in the morning). Based on previous viewing history, clustering techniques can be used to build a ‘family profile’ to differentiate potential program sets that might be watched on a particular television set. After different sets have been identified a probability score can be assigned to all sets that would be used to predict which profile would be in use at any given time for recommendation. This solution has a drawback that efficient feedback system cannot be integrated into the system since users cannot be identified. It is also possible to create a hybrid system that would use different strategies to offer recommendation depending on whether the user is logged in or not.

3.5 Feedback and evaluation

The efficiency of the system can be evaluated by direct feedback from users. Users can rate whether the recommended programs are interesting to them. Alternatively, users can provide implicit feedback by simply watching recommended movies (Yehuda 2008). This is based on an assumption that if the user watches a multimedia content to some length, that item is interesting to them. In the digital television domain, feedback and evaluation of the system would depend how users are identified, if the program clustering technique is used to predicted who is watching the television set, it is more problematic to implement a reliable feedback system, since the feedback would be based on the same prediction that determines who is watching the program, it is questionable whether that prediction could be used to determine who gives the feedback.

4. CONCLUSION

Recommender systems made significant progress over the years. However, the current generation of recommender systems still require further improvements. Since, recommender systems applied to television still in its infancy, there are many challenges that researchers face. This paper identified the main areas of my research in the domain of personalised television. It also considered possible solutions to the anticipated problems. It is important to note that these problems are defined vaguely only considering the bigger picture. Therefore, future work should be done to understand the problem space in order to provide efficient solutions to the identified problems.

REFERENCES.

[1] Adomavicius, G., Sankaranarayanan, R., Sen, S. & Tuzhilin, A. (2005) Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS)*, Vol. 23 (1) pp. 103-145.

[2] Adomavicius, G. & Tuzhilin, A. (2005) Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Ieee Transactions on Knowledge and Data Engineering*, Vol. 17 (6) pp. 734-749.

[3] Anderson, C. (2008) *The long tail: Why the future of business is selling less of more.* Hyperion.

[4] Boyer, A. & Brun, A. (2007) Natural language processing for usage based indexing of web resources. IN Amati, G., Carpineto, C. & Romano, G. (Eds.) *29th European Conference in Information Retrieval Research (ECIR 2007).* Rome, ITALY.

[5] Chang, E. Y. (2008) Organizing multimedia data socially. *Proceedings of the International Conference on Content-based Image and Video Retrieval.* ACM New York, NY, USA.

[6] Golder, S. A. & Huberman, B. A. (2005) *The structure of collaborative tagging systems.* Information Dynamics Lab, HP Labs. Available from: http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf [accessed 24/04/2009]

[7] Herlocker, J. L., Konstan, J. A., Borchers, A. & Riedl, J. (1999) An algorithmic framework for performing collaborative filtering. *Sigir'99: Proceedings of 22nd International Conference on Research and Development in Information Retrieval*, Vol. pp. 230-237.

[8] Lee, T. Q., Park, Y. & Park, Y. T. (2009) An empirical study on effectiveness of temporal information as implicit ratings. *Expert Systems with Applications*, Vol. 36 (2) pp. 1315-1321.

[9] Sarwar, B., Karypis, G., Konstan, J. & Reidl, J. (2001) Item-based collaborative filtering recommendation algorithms. ACM New York, NY, USA.

[10] Schein, A. I., Popescul, A., Ungar, L. H. & Pennock, D. M. (2002) Methods and metrics for cold-start recommendations. ACM New York, NY, USA.

[11] Tsunoda, T. & Hoshino, M. (2008) Automatic metadata expansion and indirect collaborative filtering for TV program recommendation system. *Multimedia Tools and Applications*, Vol. 36 (1-2) pp. 37-54.

[12] Wang, J., De Vries, A. P. & Reinders, M. J. T. (2006) Unifying user-based and item-based collaborative filtering approaches by similarity fusion. ACM New York, NY, USA.

[13] Wang, J., Pouwelse, J., Fokker, J., De Vries, A. P. & Reinders, M. J. T. (2008) Personalization on a peer-to-peer television system. *Multimedia Tools and Applications*, Vol. 36 (1-2) pp. 89-113.

[14] Yehuda, K. (2008) Factorization meets the neighborhood: a multifaceted collaborative filtering model. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* Las Vegas, Nevada, USA, ACM.

Supporting Polyrepresentation and Information Seeking Strategies

Thomas Beckers
Department of Computer Science and Cognitive Science
Universität Duisburg-Essen
47048 Duisburg, Germany
http://www.is.inf.uni-due.de
tbeckers@is.inf.uni-due.de

Abstract

This paper introduces the basic concepts and notions of a new framework for interactive information retrieval. Based on examples for a real-life collection of book data we show why current systems are not sufficient. It is necessary to support both polyrepresentation of information objects and multiple information seeking strategies in order to cope with the shortcomings of most current retrieval systems. A search operator concept is introduced which controls the process of retrieval. We provide real-life examples of how information seeking strategies can be supported. Furthermore, we show why interaction should play a more important role than it does today. Finally, we give an outlook about our future plans and upcoming research challenges.

Keywords: interactive information retrieval, polyrepresentation, information seeking strategies

1. INTRODUCTION AND MOTIVATION

Most existing information retrieval systems support a very limited view on documents. Usually, only the content of a document is regarded, which is made searchable by using a single representation. Furthermore, in the majority of cases there is just limited support for interaction between the user and the system.

Based on an analysis of information retrieval as an information seeking activity, Belkin [1] suggests that interactions with documents should play an important role during the retrieval process. He defines a set of information seeking strategies (ISSs) and points out that different ISSs require different interactions and search behaviours. Thus, traditional user interfaces and systems that are solely based on simple representations of texts and assume well-defined queries cannot satisfy all possible information needs of the user.

In this paper, we propose a concept that deals with the aforementioned challenges and that aims at improving the retrieval process by incorporating different aspects of documents and supporting a variety of information seeking strategies.

As an application example for demonstrating our ideas, we regard Amazon's[1] book catalog in combination with LibraryThing's[2] book metadata. At the moment, it is not possible to search in Amazon e. g. for the content of books or front covers and neither for data about the dimensions nor for the number of pages of books. Although searching in customer reviews is possible in principle, it is very restricted and deeply hidden inside Amazon[3]. LibraryThing provides even less possibilities for book search. In both systems the interaction is restricted to defining the search query and sorting the result list according to certain attributes like e. g. publication date or price. Both search interfaces can be considered as exemplary for many current real-world systems.

[1] `http://www.amazon.com`

[2] LibraryThing's is a social book cataloging web application which allows its user to store and share their personal library catalogs, `http.//www.librarything.com`

[3] There is a text field at the review detail page of a book. If the checkbox is disabled a search can be performed on all reviews. All statements concerning Amazon and LibraryThing are as of April 2009.

For demonstrating our ideas, we built a new test collection by crawling the meta-data of approx. 2.7 million books from Amazon as well as from LibraryThing and merged them into a coherent structure. The data from Amazon includes the creators (e. g. author, editor, illustrator), title, publisher, dimensions (height, width, length, weight), classifications (reading level, subjects, browse nodes), thumbnails of the cover, similar products, editorial and customer reviews. LibraryThing supplies user generated data about blurbers[4], dedications, epigraphs, first words, last words, quotations, series, awards, people, places and tags.

The aim of our proposed framework is to expand classic retrieval systems by allowing the user to interactively use rich representations of documents for retrieval while as many ISSs as possible are supported.

2. RELATED WORK

Ingwersen proposed the concept of *polyrepresentation* as a general framework for interactive information retrieval [2]. All participating cognitive structures are of potential value on both the system side and the user side. To support different cognitive structures as many and as different representations of information objects (documents) and information needs as possible should be used for retrieval. This principle of intentional redundancy is called polyrepresentation. Ingwersen's work supports our hypothesis that single, simple representations of documents are not sufficient to allow effective information retrieval.

Belkin et al. [3] propose four facets for classifying ISSs. In this paper, we focus on the three facets *method*, *mode* and *goal* of seeking. The method can be either *searching* or *scanning*, whereas searching refers to a targeted and focused search while scanning is the mostly sequential examination of a result list. The mode can either be *specification* if the user is able to express his information need or *recognition* if he is not. The goal can either be *learning* about e. g. documents or *selection* of documents. A first approach for a system supporting multiple ISSs is described by Yuan and Belkin [4], which focuses on the *method* facet.

In contrast to the traditional understanding of computation the paradigm of *interactive computation* [5] does not reduce every task to a simple function but regards interaction as an important part of solving a task. Wegner [6] and others claim and circumstantiate that interaction can be more powerful than classic algorithms. Brought forward to the domain of information retrieval, interaction should play a bigger role. Currently, most research still focuses on the query formulation – result computation cycle and aims at optimising the latter. Instead, we should allow for richer interaction possibilities, making interactive retrieval more flexible such that it can adapt to the different ISSs.

3. POLYREPRESENTATION, SEARCH OPERATORS AND INTERACTION

We think that there are three important concepts which are essential for effective interactive information retrieval supporting different ISSs: polyrepresentation, search operators and interaction, which we discuss in the following.

3.1. Polyrepresentation

Our notion of polyrepresentation is broader than Ingwersen's, by comprising as many aspects of information objects as possible. Fig. 1 shows an example of our interpretation of this concept. Every facet of a *document* can be modelled by a so-called *aspect*. For each aspect, there may be various *representations*, which form the reference points for searches. In our opinion, this broad notion of polyrepresentation is more adequate for the type of information objects we are dealing with today.

The Lacostir project [7] suggests three aspects, namely layout, content and structure. The content is what classic information retrieval is about. The structure aspect comprises the logical structure of documents as it is common for database-oriented XML retrieval like e. g. in XQuery without full-text search; classical examples of this aspect are document attributes like e. g. author or publication year. The layout aspect is concerned with the displaying of documents on mediums. These aspects can serve

[4] A blurber is someone who writes a very short – mostly positive – review for the back cover of books.

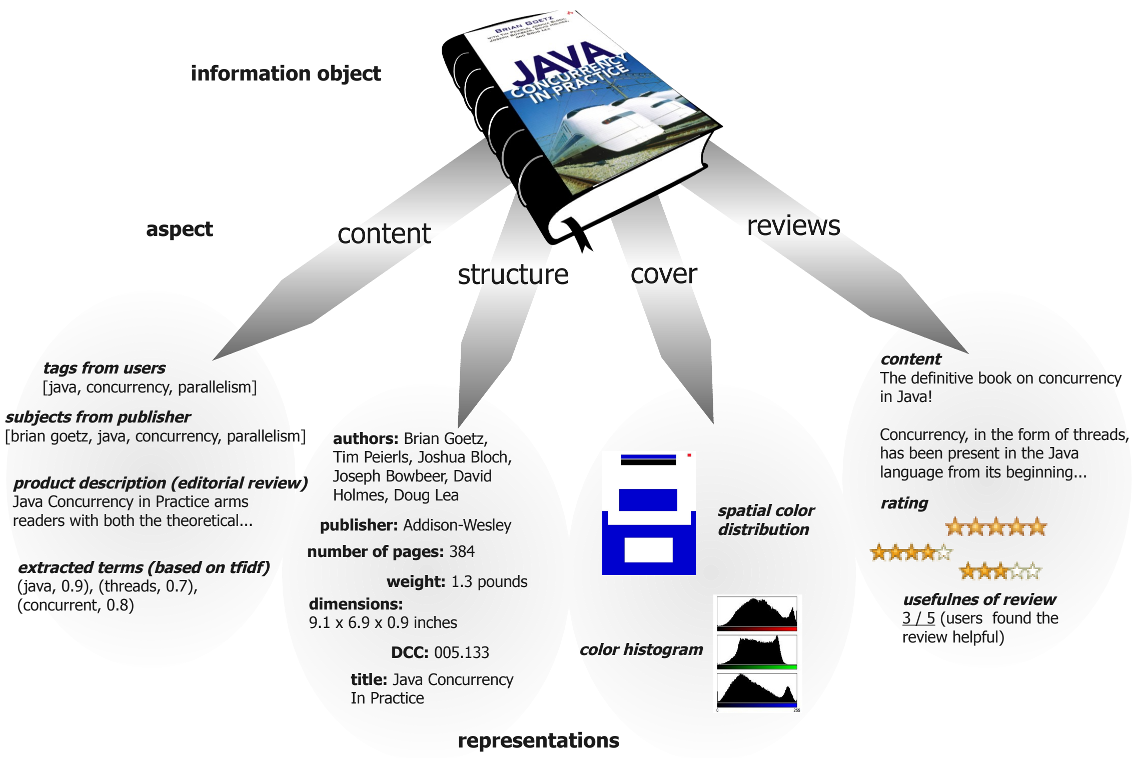

information object
aspect
content
structure
cover
reviews
tags from users
[java, concurrency, parallelism]
subjects from publisher
[brian goetz, java, concurrency, parallelism]
product description (editorial review)
Java Concurrency in Practice arms readers with both the theoretical...
extracted terms (based on tfidf)
(java, 0.9), (threads, 0.7), (concurrent, 0.8)
authors: Brian Goetz, Tim Peierls, Joshua Bloch, Joseph Bowbeer, David Holmes, Doug Lea
publisher: Addison-Wesley
number of pages: 384
weight: 1.3 pounds
dimensions: 9.1 x 6.9 x 0.9 inches
DCC: 005.133
title: Java Concurrency In Practice
spatial color distribution
color histogram
content
The definitive book on concurrency in Java!
Concurrency, in the form of threads, has been present in the Java language from its beginning...
rating
usefulnes of review
3 / 5 (users found the review helpful)
representations

FIGURE 1: Our notion of polyrepresentation of information objects described by a sample book of our collection from Amazon/LibraryThing

as a generic polyrepresentation that fit to nearly all possible documents. Depending on the concrete application these aspects can be redefined or new aspects and representations can be added.

In Fig. 1, a possible definition of aspects for our Amazon/LibraryThing book collection is depicted. The content aspect incorporates representations of the actual content of a book, such as tags or editorial reviews. Representations of the structure (e. g. author, publisher) are part of the structure aspect. Several representations of thumbnail images of a book cover belong to the cover aspect, like e. g. color histograms or spatial color distributions. Since Amazon allows its users to write reviews about books, the aspect reviews contains all representations of reviews such as the actual content or the rating ranging from one up to five stars.

3.2. Search Operators

Our *SOPV* model for interactive retrieval consists of four steps, namely *selection*, *organisation*, *projection* and *visualisation*, hence its name. We call operations in each of these steps *search operators*. These operators model the interactive process of information retrieval. The steps of this model correspond to the reference model for information visualisation [8] proposed by Shneiderman. Due to polyrepresentation of documents with respect to their aspects, different operators are essential for different representations of documents. Following, the four steps and some possible operators are described in some more detail.

Selection A user formulates selection conditions that pick possible relevant documents from one or more document collections. This step includes the choice of search queries, retrieval models and document collections. It is possible to search with respect to the various aspects but of course only on available representations (e. g. searching at Amazon for a book showing a high-speed train on the front cover is not possible).

Organisation The selected documents can be organised in various ways. Possible forms of organisation are the traditional list sorted by retrieval status value as it is used by most current retrieval systems or clustering the results by similarity of certain representations. The choice of the organisation also depends on the choice of aspects. Sorting documents according to the content is not possible while sorting after certain structure aspects (e. g. publication date) is reasonable. Consider a thumbnail of a book cover and tags describing the book's content. Performing clustering on those representations would require a variety of appropriate clustering operators [9]. So, the set of possible operations depends on the types of the representations that are used for organisation. Summing up, a user can e. g.

- organise the selected documents as list or as a 2/3-dimensional table or space, sorted by certain attributes or
- perform clustering with regard to some representations or attributes thereof.

Projection The user may be only interested in certain attributes of representations, e. g. the title and the authors of the structure aspect. This filtering is applicable on structure attributes while other types of representations should allow for different projections. A possible projection for content representations (e. g. the content of reviews) are query-based summaries which project the content to a short summary related to the search query (see e. g. [10]). As described in the two examples above the choice of projection operations depends on the type of the representation.

Visualisation Finally, the selected, organised and projected representations are visualised. This visualisation step is required given that there are countless possible visualisations for a single visual structure which is the result of the steps performed before.

This model can be described best by means of an example based on our book collection (see fig. 2). A user needs information about concurrency in Java. He heard about a good book about concurrency in Java with an express train on the front cover some months ago. He can't remember the name of the author but he would be able to recognise him if he reads his name. He *selects* books that match his query `java concurrency`. Then, he *organises* the resulting books as a list ordered by the retrieval status value whereas he *projects* only on attributes he's interested in: authors, title and a thumbnail of the front cover. Finally, the user chooses a *visualisation* that is similar to that of Amazon. The first two books seem to be relevant. He's now able to recall the name of the author, namely Brian Goetz. So, the first book is the one he is searching for.

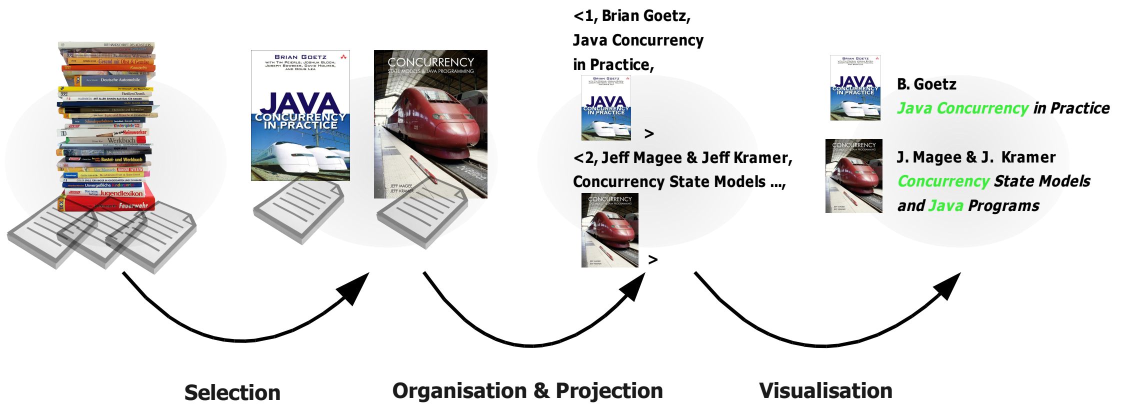


FIGURE 2: Example of the SOPV model based on our book collection

3.3. Interaction

The SOPV model offers many possibilities for interactions: the choice and configuration of the three search operators as well as interaction with the visualisation. Classic interaction techniques include panning & zooming, focus + context and highlighting (e. g. of terms contained in the search query). Also, more advanced interaction techniques are possible, such as query by example by allowing the user to specify search and projection operations through exemplary marking of certain attributes of representations: Given an example document, the user could edit some of its attributes and specify them as search conditions [11]; furthermore, by highlighting certain parts of the entry, the user could indicate that only these parts should be shown for each result item.

4. SUPPORTING INFORMATION SEEKING STRATEGIES

The various ISSs can be supported by different combinations and configurations of search operators and interaction. There are many possible types of different ISSs. Following, some examples of ISSs referring to the aspects are outlined.

(i) I am looking for a cookbook about the Chinese cuisine that has good reviews. Thus, I start a search for books about `chinese cuisine`. Only books with a rating greater than 4 stars should be shown. The result list should be sorted by the average rating (searching, specification, select / content and review aspects)

(ii) I need a book as a gift for my girlfriend. I only know for sure that her favourite books are novels about a police inspector in a Scandinavian country. So, I want to cluster novels by their content. The most important terms should also be shown in order to recognise relevant books. (scanning, recognition, select / content aspect)

(iii) A friend of mine has a guidebook about New Zealand that I want to use during my next holidays. I want to know if it covers all places which I want to visit. Therefore I search for a guidebook via authors and title to learn if a summary of this book's content contains terms like e. g. `milford sound` (searching, specification, learning / content aspect)

For supporting (i) the user may want to project only on attributes, like e. g. the author and the title summary, that are well suited for identifying the relevant books while supporting (ii) the user may decide to use additional projections, e. g. a query-based summary of the book's content, and a different organisation, e. g. a list sorted by publication date. If he can remember the front cover (cover aspect), he would probably apply an organisation operator which clusters books by similarity of their front covers.

Amazon only offers limited support for ISSs. Searching as method of seeking is only possible based on the structure aspect of books. One can organise and project according to e. g. the publication date while searching in reviews or cover images is not possible. Amazon does not offer operations on the cover and review aspect because there are no representations for these aspects (i) to create appropriate surrogates of documents that allow searching. While searching is not well supported on these aspects, scanning on the cover aspect is possible since there are thumbnails of them allowing scanning. However, scanning for content (ii) or review aspects is impossible because there are no operations to create adequate document surrogates. Learning as goal of seeking (iii) as well as recognition as mode of seeking (ii) is not directly supported at all.

Overall, we see that it is technically possible to create polyrepresentations covering all document aspects, which can be used for defining appropriate search operators. Thus, we want to develop an interactive framework, which in turn forms the basis for supporting ISSs.

5. USER INTERFACES FOR INTERACTIVE RETRIEVAL

Due to the variety of polyrepresentation and ISSs different or flexible configurable user interfaces are required. For instance, an ISS that needs to visualise clusters of documents needed for scanning or recognition must be treated with a different user interface than an ISS that relies on the retrieval of known items. However, the interface of most search engines (including Amazon's) are of rather static nature. The design of the optimal user interface is subject of our future research. One main challenge is to find a good balance between flexibility and complexity. A highly flexible user interface supporting many ISSs and polyrepresentation would allow maximum search and interaction possibilities but the increased complexity may confuse or overwhelm in particular unexperienced users and at the worst even experienced users like e. g. librarians.

As described above, the underlying principles of classic user interfaces for retrieval are not satisfactory, thus we aim at developing a concept for flexible yet easily usable user interfaces that support our concepts. We want to find out how an optimal user interface for interactive retrieval that supports ISSs, polyrepresentation and interaction should look like.

6. CONCLUSION AND OUTLOOK

In this paper we have outlined a new approach for interactive information retrieval. The concepts of rich and diverse polyrepresentations of documents as well as a model based on selection, organisation, projection and visualisation to support various ISSs were provided. We have laid out the basic notions that are required for effective interactive information retrieval. We have illustrated our concepts using a collection of book data.

Future research questions include whether our framework is actually adequate for effective interactive retrieval and if the SPOV model can be incorporated into a more holistic framework. It is intended to carry out a student project that aims at developing innovative user interfaces for searching in our book data. Currently, we are planning to do first implementations of our ideas based on Daffodil[5]. Therewith, we aim at performing evaluations at the iTrack of the INEX evaluation initiative [6] with our book data.

REFERENCES

[1] Nicholas J. Belkin. Interaction with texts: Information retrieval as information seeking behavior. In *Information Retrieval '93.*, pages 55–66, Konstanz, 1993.

[2] P. Ingwersen. Polyrepresentation of information needs and semantic entities, elements of a cognitive theory for information retrieval interaction. In *Proc. of the 17th Annual International ACM SIGIR Conference*, pages 101–111, 1994.

[3] Nicholas J. Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9:1–30, November 1995.

[4] Xiaojun Yuan and Nicholas J. Belkin. Supporting multiple information-seeking strategies in a single system framework. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference*, pages 247–254, New York, NY, USA, 2007. ACM.

[5] Dina Q. Goldin, Scott A. Smolka, and Peter Wegner. *Interactive Computation. The New Paradigm.* Springer, 2006.

[6] Peter Wegner. Why interaction is more powerful than algorithms. *Co. ACM*, 40(5):80–91, 1997.

[7] Norbert Fuhr, Matthias Jordan, and Ingo Frommholz. Combining cognitive and system-oriented approaches for designing IR user interfaces. In *Proc. of the 2nd International Workshop on Adaptive Information Retrieval (AIR 2008)*, October 2008.

[8] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors. *Readings in information visualization: using vision to think.* Morgan Kaufmann, San Francisco, CA, USA, 1999.

[9] Kerry Rodden, Wojciech Basalaj, David Sinclair, and Kenneth Wood. Does organisation by similarity assist image browsing? pages 190–197, 2001.

[10] Hideo Joho and Joemon M. Jose. Effectiveness of additional representations for the search result presentation on the web. *Inf. Process. Manage.*, 44(1):226–241, 2008.

[11] K. Großjohann, N. Fuhr, D. Effing, and S. Kriewel. Query formulation and result visualization for XML retrieval. In *Proceedings ACM SIGIR 2002 Workshop on XML and IR*, 2002.

[5] `http://www.is.inf.uni-due.de/projects/daffodil/`

[6] `http://www.inex.otago.ac.nz/tracks/interactive/interactive.asp`

Context-aware retrieval going social

Luca Vassena
University of Udine - Dep. of Mathematics and Computer Science
Via delle Scienze 206,
33100 Udine, Italy
luca.vassena@dimi.uniud.it

Abstract

In this paper we present the Social Context-Aware Browser, a general purpose solution to Web content perusal by means of mobile devices. This is not just a new kind of application, but it is a novel approach for the information access based on the users' context. With the aim of overtaking the limits in current approaches to context-awareness, our solution exploits the collaborative efforts of the whole community of users to control and manage contextual knowledge, related both to situations and resources. This paper presents a general survey of our solution, describing the idea and some scenarios, presenting the model to information access, open problems and future challenges.

Keywords: context-aware retrieval, mobile search, social, folksonomy, evaluation

1. INTRODUCTION

The widespread diffusion of mobile devices and, with them, of real-world mobile users, have moved the static world of classical and Web IR towards an always changing context-based world. Dynamism and evolving situations have become the central elements of the environment where information retrieval is asked to operate. The dynamic nature of the user needs, of the information available, and of the relevance of this information, call for new approaches to application development, user-device interaction, and information seeking. So, the notion of *context* (roughly described as the situation the user is in), and the information it conveys, are gaining increasing importance for the development of new IR systems. When combined with context-awareness, IR has been named Context-Aware Retrieval (CAR) [4].

We can imagine a user seeking information on the Web. In a traditional situation, she has to manually interact with search engines, making explicit her information need into a query and filtering out the not relevant retrieved resources. If this can be acceptable in the everyday use of a desktop system, it becomes a serious issue when the same task has to be carried out in a mobile environment. These considerations guided us towards a new approach to Web contents production and use named *Social Context-Aware Browser*. The novelty of the proposed approach is threefold. First of all this is a new radical approach that aims at discovering *"the query behind the context"*: to retrieve what the user needs, even if she did not issue any query [11]. Second this is not a domain dependent application, but a new generic way of interaction and information access, able to adapt to every domain. Third, as current models for context-awareness are too limited for very general applications, this approach brings new models for CAR that exploit the collaborative efforts of the community of users.

In this paper we first briefly survey CAR system (Section 2). In Section 3 we present the main idea of our approach, introducing some scenarios, and underlying how current models for contextual knowledge management are not suitable for our solution. We then present our new model for the information access based on context (Section 4): we present the conceptual model and the open questions, focusing on possible methods to evaluate the model. At the end, in Section 5 we draw some conclusions and present future work.

2. CONTEXT-AWARE RETRIEVAL

Context-Aware Retrieval (CAR) is an extension of classical Information Retrieval (IR) that incorporates the contextual information into the retrieval process, with the aim of delivering information to the users that is relevant within their current context [7]. CAR systems are concerned with the acquisition of context, its understanding, and the application of behaviour based on the recognized context [15]. Thus the CAR model includes, among the classical IR model elements, the user's context, that is both used in the query formulation process and associated with the documents that are candidates for retrieval.

Typical CAR applications present the following characteristics [7]: a mobile user, i.e., a user whose context is changing; interactive or automatic actions, if there is no need to consult the user; time dependency, since the context may change; appropriateness and safety to disturb the user. Although CAR applications can be both interactive and proactive in their communication with the user, we concentrate on the proactive aspects, since they are more relevant to our proposal. Besides, we concentrate on the association between CAR and mobile application, as they can be considered as the prime field for CAR [7].

A first exploitation of user's contextual information for the retrieval of documents is represented by the idea of "virtual Post-It" [3, 5]. More advanced examples are CoolTown [8], AmbieSense [13], Physical Mobile Interaction [2], where mobile devices exploit contexts to extract information and services associated with physical entities. An extension of the previous approaches is represented by the Ubiquitous Web [9], that is based on the spontaneous annotation by a community of users of objects, places, and other people with Web accessible content and services. A more general system is represented by the MoBe framework [11]. In this application, a general inferential framework (based on ontologies and Bayesian networks) combines the information coming from sensors to infer new and more abstract contexts (user activities, needs, etc.), that are used to retrieve and execute the most relevant applications.

3. SOCIAL CONTEXT-AWARE BROWSER

3.1. Description

The Social Context Aware Browser (sCAB for short) is a general purpose solution to Web content navigation by means of context-aware mobile devices. It allows a "physical browsing": browsing the digital world based on the situations in the real world. The main idea behind sCAB is to empower a generic mobile device with a browser able to automatically and dynamically retrieve and load Web pages, services, and applications according to the user's current context.

The sCAB acquires information related to the user and the surrounding environment, by means of sensors installed on the device or through external servers. This information, combined with the user's personal history and the community behaviour, is exploited to infer the user's current context (and its likelihood). In the subsequent retrieval process, a query is automatically built and sent to an external search engine, in order to find the most suitable Web pages for the sensed context and present them to the user.

Considering the sCAB usage we can find four main scenarios of interaction, where values from sensors and resources candidates for retrieval can be enhanced with contextual information. In the first case, directly on the basis of the information provided by sensors, a resource is retrieved. For example, inside a museum, the sCAB perceives the wireless network named *Museum X* and retrieves the Web page whose content is the presentation of that museum. With a subsequent step, we annotate resources with contextual information, in order to retrieve them in the right situation. For example a Web page describing an historical fact can be enhanced with location information (GPS coordinates), to be automatically retrieved only when users are in that place. In the same way we can enhance the knowledge related to contexts, making an abstraction on the sensors' values. For example information about activities can be related to a particular combination of sensors' values, in order to retrieve resources based on the activity the user is doing: when the user is taking the dog out for a walk, Web pages that speak about dogs training are retrieved. In the last and more general step, we can imagine a user in a museum: when she is near an artwork, a detailed description is presented on her device. In a crowded situation, on the contrary, a detailed description is not useful as users can have difficulties in seeing the paintings, thus a navigable high resolution picture can be more interesting. In this case, both resources and sensors' values are enhanced with contextual information.

In a such general and large scale approach, contextual knowledge is continuously associated (added, removed, and modified) with resources and low level sensors information. This entails the creation and management of a huge amount of knowledge related to contexts. Thus, the main question is to understand who is the provider of that knowledge and how it has to be defined.

3.2. Social approach

In current approaches to context-awareness, that manage several dimensions of context, the knowledge is usually provided by a small group of experts (application developers or specific domain experts). This is due to the difficulties in representing contexts. In fact, in order to fully capture the concept of context, these approaches are based on categorizations and ontologies, and implicates the strict definition of the contextual information. Moreover contexts are defined a priori, and there is no way to dynamically extend the contextual values adopted or to enhance their representation at run-time: the operations of modeling contexts and using context-aware applications are rigidly separated. This is the reason why current approaches show a trade off between the generality of applications and the depth of context representations. Applications that fully manage several contextual dimensions are confined to limited fields (e.g. Smart Homes), while general applications work only on a narrow notion of context (e.g. in location-based applications the context is represented just by location and time).

The high generality and the deep context representation we aim at with the sCAB, require both a dynamic nature and a huge amount of information to be categorized and modeled (to represent both contexts and contexts-resources associations). For these reasons, current approaches are not suitable for the sCAB.

Starting from these considerations, we propose a novel model for CAR, that aims at overtaking the just defined limits, exploiting the social dynamics underlying the Web 2.0. In fact we believe that only the collaborative effort of a community can provide the right tool for a comprehensive definition, management and use of context, in an open architecture as the sCAB. In particular, we do not want a priori contexts definitions made by experts, and we do not want people to be just passive users. Rather, through collaborative annotation, the community of users is encouraged to define the contexts of interest, share, use and discuss them, associate context to content (web page, applications, etc.), to have a dynamic and more user-tailored context representation and to enhance the process of retrieval based on users' actual situation.

4. PROPOSED APPROACH

4.1. Conceptual model

Users, contexts and resources are the main elements of our model (Fig. 1). A *user*, in the real world, is engaged in some activities, she has some needs, and she perceives her surrounding environment through her senses and the sensors on her mobile device. *Contexts* are virtual representations of the users current situations. *Resources* represent every kind of content that could be useful for the user to accomplish her needs.

Instead of using rigid categorizations built upon ontologies and terminologies, we represent the context by means of a folksonomy. This representation allows an easy and informal modelling of context, giving the opportunity also to non-expert users to classify and find context-related information. Thus each context is represented by a tag cloud and the tags (little squares in Fig. 1) are socially *defined* by the users themselves. The tag-context association is done both explicitly, when the tags are added directly by the user, and implicitly, when the tags are derived from the interaction of the community with resources.

Six are the main operations in the model [12].

Contexts definition: users can explicitly use tags to represent contextual information. For example, the user can enhance the values provided by sensors on the mobile device ("concrete" values) with her own tags, as `''out dog walk leash park play ball''`. Doing so these "abstract" tags are stored in a remote repository and they are linked with the concrete ones. For all the users with the same or a similar concrete tag cloud, the abstract tags (or part of them) can be retrieved and become part of the representation of their context.

Contexts inference: the values provided by sensors are combined with the contextual tags defined by the whole community, and tags that best describes the user's situation are retrieved. For example, starting from the GPS coordinates, the current user's context could be enhanced with the tags `''walk sunny park''`.

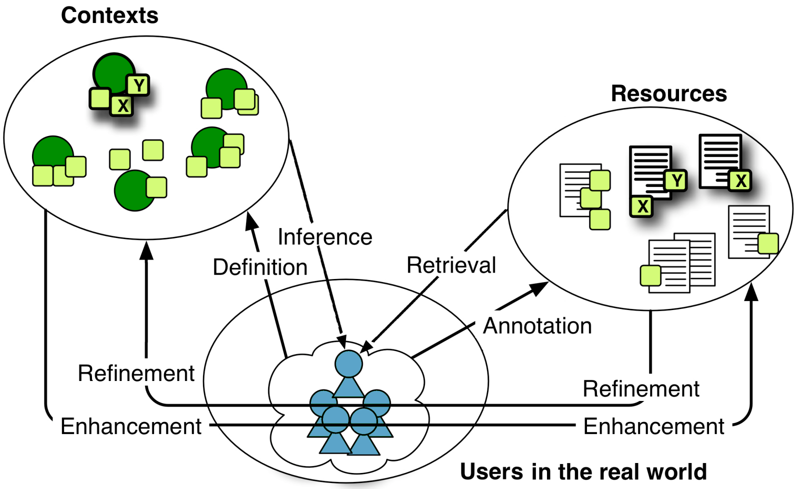


FIGURE 1: A conceptual model for social CAR

Resources annotation: users can explicitly annotate resources with contextual information to allow the community of users to automatically retrieve them when they are in the suggested context. For example a user can associate a classical music web-radio with the context ``out dog walk'', in order to listen to music when she is out with her dog.

Contextual retrieval: based on users' current context and on the contextual information associated to resources, the most relevant resources are retrieved.

Refinement: information about contexts are refined based on the interaction between users and resources. For example, if a lot of users work with a Web application in the same context ``work'', probably this resource is related to that context and it is automatically annotated with it.

Enhancement: information related to resources are refined based on the interaction of users within their contexts. For example if a user uses resources annotated with the context ``work'', probably she is working, and the representation of her context can be enhanced with this information.

Although the knowledge related to the whole community is exploited to infer and refine the current context of single users, the proposed model differentiates the personal from the community level, giving more importance to the first one. For example if a user annotates a situation as ``play'', she is considered to be in ``play'' context, even if most people annotate the same situation as ``work''. On the contrary, if a user is for the first time in a situation (e.g. location never visited), her context is refined just with the information from the community. Considering the previous example, as most people annotate the situation with ``work'', the user is considered to be in ``work'' context.

4.2. Issues and open questions

Several are the issues related to the proposed approach: they go from implementation/system/architecture issues, to telecommunications issues, from annotation/Web 2.0 issues, to interface/HCI, to evaluation issues. The proposed work do not address all these, but focuses on the annotation, Web 2.0 and evaluation issues. The main goal is to improve the effectiveness of the Context-Aware Browser, supporting and enabling the effective and efficient access to relevant information. Thus we want to understand if the exploitation of the crowdsourcing is viable for the management of the concept of context in a general approach as the sCAB.

In particular several critical questions can be found in the presented approach. Starting from a low level, the questions are related for example to context representation: it is better to exploit a simple folksonomy or to introduce a level of complexity (e.g. logic operator over tags)? Tagging a tag can introduce useful information? Tags should be associated with probability values to describe the uncertainty related to contexts?

Other interrogative points are related to the association between tags and the elements in the model. New tags are continuously added to situations and resources, leading to a huge amount of tags for each of them. Thus the question is how to understand, given a resource or situation, which are the most relevant contextual tag. All, just the last ones used, or have we to consider the number of times a tag has been related to a resource? Or rather we can imagine a more complex approach based on social evaluation, where every association (e.g. tag–resource) has a score that increases or decreases based on the community behaviour and on how users are representative for the community. Could this approach improve the relevance computation or does it introduce a useless layer of complication?

Once we have identified relevant contextual tags, which is the best strategy to combine the user's context tags with the ones associated to resources, to retrieve the most relevant resources for the given context? How can the system extrapolate contextual knowledge from the simple amount of tags provided by users? Can current machine learning and artificial intelligence techniques cope with this problem? As the users are the main actors in the process of contexts definition, how can we ensure the quality of the information provided? Moreover how can we easily manage the problem of communities and subcommunities?

Finally there are some general questions. This vision presents an extension of the idea of Web, where resources and documents are indexed not only based on their content, but also based on the context which they are relevant to. Could this model be helpful in understanding the relationships between context and content, and how do they reflect one to each other? Which is the connection between context and information needs? Is the context alone enough to understand what the user needs and how she needs it (textual information, audio, image)?

4.3. System evaluation

Although the conceptual model is clear, several alternative strategies to the problems presented exist, so it is important to compare their effectiveness. Moreover the understanding of the good and weak points of the proposed model is the first step toward its realization.

4.3.1. Evaluation approaches

Evaluation is an issue of paramount importance in IR. Depending on resources and aims, different evaluation approaches can be adopted; benchmark-based and user-centred are the main ones. The benchmark evaluation (e.g. TREC initiative, `http://trec.nist.gov/`) follows the Cranfield model, giving emphasis on controlled laboratory tests, without user interaction. Benchmarks are system centred and they directly focus on the evaluation of different implementation strategies. Within the CAR field, benchmarks have been for example exploited to have a rough evaluation during the first stage of development [10, 11].

While benchmarks concentrate on details, user evaluation approach works on the IR system from a much broader point of view. The main purpose is to study usability and interaction, to understand how the system satisfy the user's needs, and, in general, how well the user, the retrieval mechanism, and the database interact extracting information under real-life operational conditions [1]. Within the CAR field, this kind of evaluation has been for example exploited in [6, 16].

In the last years hybrid evaluation models have been studied to combine the advantages of both the previously presented approaches. These ones are mainly addressed to improve the evaluation in the interactive information retrieval field (IIR). The overall purpose is to facilitate the evaluation of IIR systems as realistically as possible, taking into account the dynamic natures of information needs and relevance as well as reflects the interactive information searching and retrieval processes (as in a user-based study), though still in a relatively controlled evaluation environment [1].

4.3.2. Suggested methodology

With this considerations in mind, we propose a multistage approach, where implementation and evaluation processes will proceed hand in hand. Different stages of work will require different evaluation solutions, and we believe that early stage benchmark evaluations, followed by user studies, is an effective methodology to be applied to systems like the sCAB. We will move from pure laboratory studies, to simulated environments with users involvement, to a real world operational environment.

Benchmark evaluation will be our first step, and will be exploited to evaluate detailed implementation solutions, like, for example, different algorithms to assess the relevance of tags for situations and resources. Benchmarks do not substitute the user testing evaluation. Rather, several early stage benchmark experiments could provide more solid basis for the subsequent user testing, that can thus be more focused. For example, knowing which is the best strategy allows us to give to users just one prototype, instead of different prototypes, one for each strategy.

Once the system has attained the desired levels of accuracy and effectiveness, we can apply an IIR evaluation methodology, involving users in a controlled environments, following the ideas presented [1, 14, 16]. In particular this step will consist in an iterative process based on the design-evaluation cycle described in [14]: starting from some hypothesis, we will build/refine a prototype, that will be evaluated, and the results will be the basis for the next cycle hypothesis.

Finally a broader user-centred evaluation will help us to understand if the sCAB is effective in the real world. This last stage does not involve laboratory experiments anymore, but only studies in an operational environment. In this stage, we must move beyond performance and usability and consider utility or impact measures [16]. That is, how do the proposed system change the work that users are doing?

5. CONCLUSIONS

In this paper we have presented the Social Context-Aware Browser, a general purpose solution to Web content perusal by means of mobile devices. Introducing the main ideas, we have shown how current approaches to contextual knowledge management are unsuitable for our solution. Thus the sCAB is not merely an application, but it is a novel paradigm for the information access based on context, where the community of users is called to manage the contextual knowledge through collaboration and participation. We presented some scenarios and the conceptual model, suggesting a possible way to evaluate it.

The project is an initial stage. As future work we aim at answer all the open questions, advancing in the same time with the evaluation of precise aspects of the model and with its implementation.

REFERENCES

[1] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3):8–3, 2003.

[2] G. Broll, S. Siorpaes, E. Rukzio, M. Paolucci, J. Hamard, M. Wagner, and A. Schmidt. Supporting mobile service usage through physical mobile interaction. *IEEE Intl. Conf. on Pervasive Computing and Communications*, pages 262–271, 2007.

[3] P. J. Brown, J. D. Bovey, and C. Xian. Context-aware applications: from the laboratory to the marketplace. *IEEE Personal Communications*, 4(5):58–64, 1997.

[4] P. J. Brown and G. J. F. Jones. Context-aware retrieval: Exploring a new environment for information retrieval and information filtering. *Personal and Ubiquitous Computing*, 5(4):253–263, 2001.

[5] J. Burrell, G. K. Gay, K. Kubo, and N. Farina. Context-aware computing: A test case. In *Proc. of the 4th Intl. Conf. on Ubiquitous Computing (UbiComp '02)*, pages 1–15, London, UK, 2002. Springer-Verlag.

[6] A. Göker and H. Myrhaug. Evaluation of a mobile information system in context. *Information Processing & Management*, 44(1):39–65, 2008.

[7] G. J. F. Jones and P. J. Brown. Context-aware retrieval for ubiquitous computing environments. In *Mobile HCI Workshop on Mobile and Ubiquitous Information Access*, volume 2954, pages 227–243. Springer LNCS, 2004.

[8] T. Kindberg, J. Barton, J. Morga, G. Becker, I. Bedner, D. Caswell, P. Debaty, G. Gopal, M. Frid, V. Krishnan, H. Morri, C. Pering, J. Schettino, B. Serra, and M. Spasojevic. People, places, things: Web presence for the real world. In *3rd IEEE Workshop on Mobile Computing Systems and Applications (WMCSA 2000)*, volume 7, 2002.

[9] D. Lopez de Ipiña, J. I. Vazquez, and J. Abaitua. A context-aware mobile mash-up plaftorm for ubiquitous web. In *Proc. of 3rd IET Intl. Conf. on Intelligent Environments*, pages 116–123, 2007.

[10] S. Menegon, Davide Mizzaro, E. Nazzi, and L. Vassena. Benchmark evaluation of context-aware web search. In *Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE) in conjunction with the 31th European Conference on Information Retrieval*, 2009.

[11] S. Mizzaro, E. Nazzi, and L. Vassena. Retrieval of context-aware applications on mobile devices: how to evaluate? In *Proc. of Information Interaction in Context (IIiX '08)*, pages 65–71, 2008.

[12] S. Mizzaro, E. Nazzi, and L. Vassena. Collaborative annotation for context-aware retrieval. In *ESAIR '09: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 42–45. ACM, 2009.

[13] H. Myrhaug, N. Whitehead, A. Göker, T. E. Faegri, and T. C. Lech. Ambiesense—a system and reference architecture for personalised context-sensitive information services for mobile users. In *Proc. of the Second European Symposium on Ambient Intelligence (EUSAI '04)*, 2004.

[14] D. Petrelli. On the role of user-centred evaluation in the advancement of interactive information retrieval. *Inf. Process. Manage.*, 44(1):22–38, 2008.

[15] A. Schmidt. *Ubiquitous Computing - Computing in Context*. PhD thesis, Lancaster University, 2003.

[16] J. Scholtz. Metrics for evaluating human information interaction systems. *Interacting with Computers*, 18:507–527, 2006.

Content-based Music Access: an approach and its applications

Riccardo Miotto
Department of Information Engineering
University of Padova
Via Gradenigo, 6/B
35131 Padova (Italy)
miottori@dei.unipd.it

Abstract

At current time, the availability of large music repositories poses challenging research problems. Among all, content-based identification is gaining an increasing interest because it can provide new tools for easy access and retrieval. In this paper we describe an ongoing methodology for the content-based identification of unknown music recordings through a collection of music documents. Moreover, as future prospective scenario, identification is viewed in a more general similarity context, where also the perception of the users is considered.

Keywords: music identification, similarity, indexing

1. INTRODUCTION

Nowadays, the availability of large music collections poses challenging problems mainly related to the organization of documents according to some sense of similarity. To address this issue, from a research perspective, an interesting starting point seems to be the automatic identification of music recordings. In this context, given an unknown excerpt of a music work, the task is to retrieve all the recordings of a music collection sharing some content with the query and to deliver relevant metadata such as title, artist and other additional information.

An earlier approach to music identification was *audio fingerprinting* that consists in a content-based signature of a music recording to describe digital music even in presence of noise, distortion, and compression [4]. However, the fingerprint value is strictly related to a particular music performance, but the identification of a music work may be carried out also without linking the process to a particular recording. For example, identification of live performances may not benefit from the fingerprint of other performances, because most of the acoustic parameters may be different. The solution would be to collect all the possible live and cover versions of a music works, which is clearly unfeasible.

Thus, a good music identification approach has to be able to identify music works from the recording of a performance, yet independently from the particular performance, and to sort the elements of the collection according to some similarity measure with the query. Music works to identify may be live performances, cover versions, noisy registrations and so on.

In literature several approaches to establish whether two musical pieces share the same melodic or tonal progression have been proposed, such as in [7], [13] and [5]. Efficient and scalable systems for a cover music identification task were also proposed in [18] and [11].

Content-based identification approaches can have a direct implication in different fields such as musical rights management and licenses, learning about music, discovery new music and many other topics related with music perception and cognition. Furthermore, a general technique can be exploited in different tasks, where identification is just an application among other possible ones. At this purpose, the ongoing methodology that we proposed in [11] aims at being general enough to be useful in other tasks mainly related to the music similarity context.

Thus, in the following, Section 2 provides an overview of the methodology together with the new directions whereas Section 3 describes ideas and prospective approaches toward a more general music similarity system including also users perception.

2. A CONTENT-BASED MUSIC IDENTIFICATION ENGINE

The content-based identification approach described in this paper was firstly proposed in [10] and [11]. The objective is to identify each recordings of a score (including live performances and cover versions) through a collection of indexed high quality recordings. The assumption we made was that, if a performance is played according to the original score, it can be generally modeled and identified through the score information alone.

The system is mainly based on an application of hidden Markov models (HMMs) [15]. Since the identification through HMMs is linear with the size of the collection, an index of the collection have been built to extract from the collection a cluster of candidates to be re-ranked with the HMMs-based identification. Figure 1 provides the general structure of the system; a first prototype is also available on-line at [12].

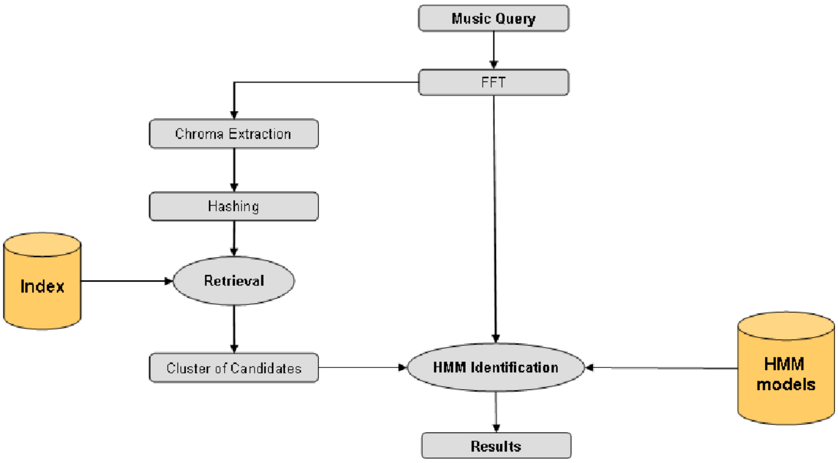


Figure 1: General Structure for the Music Identification System.

2.1. Clustering the Collection

The music features exploited to extract a cluster of the collection should be both general and robust to all the variations due to differences between the query and the collection recordings (tempo variations, tonality shift, different voices, etc.).

One of the most common music content descriptor are *chroma* features. The basic idea behind chroma is that octaves play a fundamental role in music perception and composition. For instance, the perceived quality – e.g. major, minor, dominant, and so on – of a given chord depends only marginally on the actual octaves where it spans, whereas it is strictly related to the pitch classes of its notes. Following this assumption, a number of identification techniques based on chroma have been proposed, in particular in [13] and [7].

In our approach [11], chroma features are used to index the music collection. In particular, chroma are considered as pointers to the the music documents they belong to, playing the same role of words in textual documents. Each chroma feature points to a number of recordings and to a set of time positions within each recording. One major advantage of indexing in text retrieval is that the list of index terms can be accessed in logarithmic, or even constant, time. The same cannot be applied to feature vectors, because the exact match has to be replaced by a similarity search, which is less efficient. One of the technique to handle efficiently this issue is the locality sensitive hashing (LHS) [6]. Its basic idea is to apply to the feature vectors a carefully chosen hashing function with the aim of creating collisions between vectors which are close in the high dimensional feature space. The hashing function itself becomes then an effective tool to measure the similarity between two vectors.

Following this idea, we propose to represent the 12-dimensional chroma vectors with a single integer value through an hashing function, not depending on the absolute value of the chroma pitch classes, but just on their rank within the vector. In Figure 2 the whole chroma indexing process is depicted.

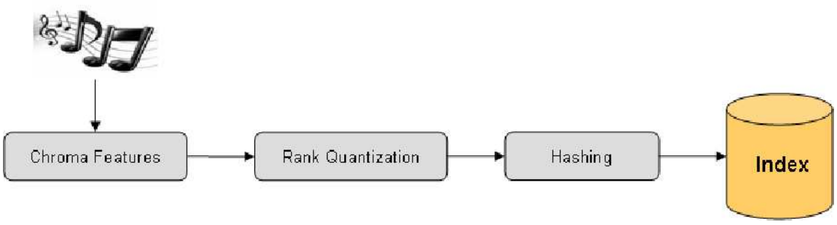


Figure 2: Chroma Indexing process, from the music recording to the index

Retrieval is carried out using the *bag of words* paradigm, then by counting the common chroma words between the query and the recordings of the collection. A problem that may affect retrieval effectiveness is that chroma-based representation is sensitive to transpositions. In fact, if the query and the matching recordings stored in the database are played in different tonalities, they have a totally different sets of chroma. We addressed this problem by considering that a transposition of s semitones will result in a rotation of the chroma vector of s steps. Each query is than transposed to include all the most common tonalities.

At the end, the top N retrieved recordings are considered as the cluster of potential candidates. Experimental results achieved with a collection of 1000 recordings of classical music showed that a cluster of 100 documents ($1/10$ of the collection) was sufficient to have a 100% of recall [11]. First evaluation with pop and rock music gave promising results even if, due to the more variety of the music, a more accurate elaboration seems to be necessary to achieve the same precision.

At this aim, experience matured in information retrieval proved that usually the combination of different features outperforms the use of just one. Then, following this assumption we believe that including other music descriptors could increase the performances of the system in terms of both precision and recall.

Considering that cover songs generally preserve not only the harmonic-melodic characteristics of the original work but also its rhythmic profile, the first idea is to combine chroma with some rhythm descriptors, such as the one proposed in [9] for a genre classification task and called *rhythm histogram* (RH). In a RH the magnitudes of each modulation frequency bin for all the critical bands of the human auditory range are summed up to form a histogram of “rhythmic energy” per modulation frequency. Similarity relationships can be measured according to the distance among the histogram representations, and the songs of the collection can be ranked following these values. A weighted final rank is then computed through a weight merge of both lists, where the greater weight is given to the chroma list.

In a similar way, another music features that could be exploited are the Mel-frequency cepstral coefficients (MFCCs). MFCCs are often used to compute music similarity especially in genre classification tasks [16, 3]. They capture the overall spectral shape of the audio signal, which carries important information about the instrumentation and its timbre, the quality of a singer’s voice, and post-production effects [2]. However, they do not capture information about melody and rhythm which are the most important features for the identification task we are proposing. Anyway, following one more time the assumption that the combination of different features usually outperforms the use of just one feature, we believe that even MFCCs can be useful to increase the efficacy of the system.

MFCCs can be represented as a sequence of n-dimensional vectors, where the value of n depends on the accuracy required by the system. Thus, they can be modeled and retrieved with the same hashing approach of chroma, and can provide another rank list of potential candidates to be merged with the other rank lists. Alternatively, considering their major application in genre classification tasks, MFCCs rank list may be exploited independently to pre-filter the collection according the music genre (pop, rock, folk, etc.). Then just the subset of the collection extracted can be considered in the chroma-based identification.

2.2. Identification

The cluster extracted from the collection is re-ranked with the HMMs-based identification methodology proposed in [10]. The basic idea of the approach is that, even if two different performances of the same music work may differ in terms of acoustic features, it is still possible to generalize their music content through a statistical models. To this aim, each recording of the collection has to pass through some modeling steps.

In a first step, a segmentation process extracts audio subsequences that have a coherent acoustic content. The aim of segmentation is to divide the music signal into subparts that are bounded by the presence of music events, where an event occurs whenever the current pattern of a music piece is modified (one or more new notes being played or stopped). Segmentation of the acoustic flow can be considered the process of highlighting audio excerpts with a stable pitch.

Coherent segments of audio are then analyzed through a second step in order to compute a set of acoustic parameters that are general enough to match different performances of the same music work. In line with the segmentation approach, also parameters extraction is based on the idea that pitch information is the most relevant information for a music identification task. Because pitch is related to the presence of peaks in the frequency representation of an audio frame, the parameter extraction step is based on the computation of local maxima in the Fourier transform of each segment, averaged over all the frames in the segment.

In a final step a HMM is automatically built to model music production as a stochastic process. The idea is that music recordings can be modeled with HMMs providing that states are labeled with events in the audio recording and their number is proportional to the number of segments, transitions model the temporal evolution of the audio recording and observations are related to the audio features previously extracted that help distinguishing different events.

At identification time, an unknown recording of a performance is preprocessed in order to extract the features modeled by the HMMs. All the models are ranked according to the probability of having generated the acoustic features of the unknown performance. Ideally the alignment of the query through the correspondent model will follow a linear trend, achieving the final higher probability. Since the simplicity of the models, coarse alignment between the events and the acoustic features could occur. This problem is handled by providing a support parameter, which measures the distance between the computed path and an estimated linear path. Such linear path can be estimated by considering the regression analysis of the computed alignment points.

A complete evaluation of the methodology with a collection composed of about 1000 recordings of classical music can be found in [11]. The complete system gave a final precision of 90% with the 84% of the analyzed query correctly ranked in the first position with an identification time of about 3 seconds. As for the clustering component, an extension for popular music is under development and initial results seems to be very promising.

3. IDENTIFICATION IN SIMILARITY ANALYSIS

A content-based music identification system may have different application scenarios, mainly towards accessing, organizing, browsing and recommending. Automatic identification of unknown recordings can be exploited as a tool for supervised manual labeling: the user is presented with a rank list of candidates, from which he can choose the matching one. Once that the unknown recording has been correctly identified, it could be indexed and added to the music collection. Identification may also be exploited to retrieve all the different versions of the query stored in the collection (live or cover).

Given the identification tool, a research prospective aims at exploiting it in a similarity measure context. At this purpose, one of the main issues concerns the understanding of the similarity concept itself. In fact, a very common consideration is about the structure of the rank list of an identification system and an immediate question may be: ”*in an ideal situation where the matching documents are always ranked on the top, is the first non-matching item the most similar of the collection to the query?*”.

The answer is not an easy task. In fact, in a mathematical sense we may say that, it is the most similar because it shares the larger number of features. However, in a general sense, the similarity concept is very subjective and strictly related to the context and especially to the listeners. Basing the similarity measures just on the content is a bit reductive because music content is very various and many factors are involved. Moreover, the perception of the users could be various and likely much different from the computed one.

In the following, we propose some solutions to exploit the identification tool in a similarity context.

Mainly, we believe that this could be done either considering the feedback with users or through an integration with textual metadata.

3.1. Users Relevance Feedback

To consider music identification as a similarity task, feedbacks provided by the users can be exploited. In the supervised manual labeling scenario previously proposed, users have to listen the documents of the rank list to find the matching items. Then, by providing a rating pool, we could measure the level of similarity with the query perceived by the listeners. In fact, people can rate all the items of the rank list according their perception of similarity (for instance with scores from 1 to 5).

Beside explicit feedback, an implicit feedback approach could also be considered. The idea is to propose the identification tool as a playlist generator. The rank list can be seen as a playlist suggested by the system in response to the query. The user could decide to listen all the retrieved items even after the matching document, for example to search new music. Then implicit feedback can be used to measure the likeness of the user to the provided playlist. Likeness can be related to songs skip, if all the song of the list were listened or not, if the song were completely listened , etc. All these measures can be processed to achieve a descriptor for the grade of likeness of the proposed playlist. This descriptor can be related to a similarity concept by assuming that the query provided is a recording that user likes and about wants to have more information. Clearly, implicit feedback must be considered just in case the user had effectively listened to the playlist, and not when he has only searched for the matching items. The time spent on the results page could be considered a valid estimator; in case of a short time, implicit measures would not be considered.

All the feedbacks could be then exploited to have similarity relations among the collection items and to create clusters of similar documents. Moreover, they could provide a study to understand the behavior of the content-based identification tool according to the similarity perception of the users (how close the ranking list is to the similarity perception of listeners).

3.2. Integration with Social Tags

The content-based identification system can be used together with textual metadata to define similarity relationships among the items of a collection. A content-based identification tool will provide a component to measure the similarity of the music content, whereas all the tags associated to the items will provide a user-based description [8].

Considering the rank list provided by the system, it would be possible to provide the system with a component to browse the collection according to some pre-computed similarity relationships based on tags. For example, starting from the matching items an user could browse the collection searching similar items or creating a playlist, where the similarity is based on social tags representing the cognitive perception of people.

A future research prospective that we are going to investigate is how to modify the structure of the identification system (Figure 1) in a tool to define off-line similarity relationships among the documents of the collection. An initial schema of the system is depicted in Figure 3.

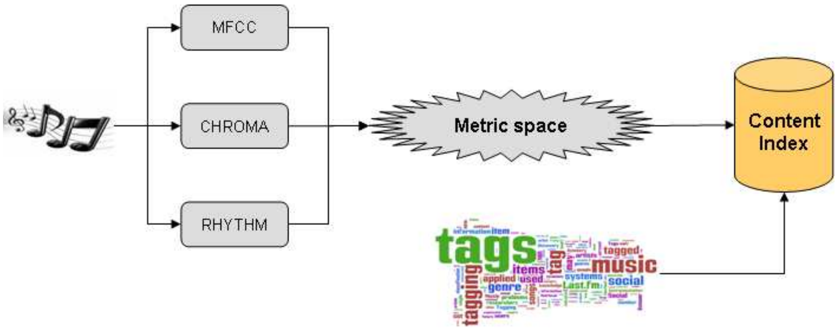


Figure 3: Schema of a system to define similarity relationships for documents of a collection.

As it can be seen, the approach aims at computing all the content descriptors and to map them in some metric spaces in order to define a sort of distance among descriptors. Considering content distances and social tags, it would be possible to define similarity relationship for the documents of the collection. We believe that these computed similarity scores would be very descriptive since representing both music content and users perception.

A preliminary idea is based on representing the collection and the similarity relationships as an hidden Markov model where transition probabilities among states depend on the content-similarity relationships whereas the observation probabilities of each state are related to the social tags. Well known, statistical paths through the model [15] allowed an user to browse the collection according a similarity global value express in terms of probability which combines both content and user similarity relationships.

All social tags can be collected in different ways. In literature different methodologies to tag music have been proposed, mainly based on either human-annotations, web mining or auto content-based annotations [14]. A common approach is to gather human-based tags for a descriptive training set and then to exploit a content-based auto-tagger approach, such as the one proposed in [16] which uses a machine learning technique based on multivariate Gaussian mixture models to annotate the new songs.

The proposed system may be used in different tasks ranging from recommendation systems to the browsing by similarity of a collection. Especially concerning recommendation, we believe that it could be an useful tool for the ”*items cold start*” issue [1], since new items will be provided with content-based descriptors that would be useful to define also likely social aspects.

4. CONCLUSIONS

This paper describes a methodology for the content-based identification of music documents. The aim is to identify each recordings of a score (including live performances and cover versions) through a collection of indexed high quality recordings. The approach is based on two steps. At first, a cluster of the collection is retrieved to highlight some potential candidates for the query, whereas a second step computes the similarity between the query and the documents of the cluster with an application of HMMs. The methodology is still under study, and prospective directions to improve the results are provided.

A content-based music identification system may have different application scenarios, where supervised manual labeling of unknown music recordings seems to be the most suitable. In this context, a larger concept of similarity, not only based on content features but also considering the cognitive perception of the listeners, can be introduced. At this aim, we proposed some different ideas to apply the music identification tool in a larger similarity context. The descriptions provided represent future approaches that we are going to investigate and we consider interesting for the community.

Bibliography

[1] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, vol.17, no.6, June 2005.

[2] A. Berenzweig, B. Logan, D. Ellis and B. Whitman. A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measure. *Computer Music Journal*, 28:2, pp.63–76, 2004

[3] T. Bertin-Mahieux, D. Eck, F. Maillet and P. Lamere. Autotagger: A Model for Predicting Social Tags from Acoustic Features on Large Music Databases *Journal of New Music Research*, vol.37(2), pp.115–135, 2008

[4] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A Review of Audio Fingerprinting. *Journal of VLSI Signal Processing*, vol.41, pp.271–284, 2005.

[5] D. Ellis and G. Poliner. Identifying Cover Songs with Chroma Features and Dynamic Programming Beat Tracking. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol.4, pp.1429–1432, 2007

[6] A. Gionis, P. Indyk, and R. Motwani. Similarity Search in High Dimensions via Hashing. In *Proceedings of the International Conference on Very Large Databases*, pp.518–529, 1999.

[7] P. Herrera, J. Serrá, E. Gómez and X. Serra. Chroma Binary Similarity and Local Alignment applied to Cover Song Identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138–1151, 2008.

[8] P. Lamere. Social Tagging and Music Information Retrieval *Journal of New Music Research*, vol.37(2), pp.101–114, 2008

[9] T. Lidy and A. Rauber. Evaluation of Feature Extractors and Psycho-acoustic Transformations for Music Genre Classification In *Proceedings of the International Conference on Music Information Retrieval*, pp.304–310, 2005.

[10] R. Miotto and N. Orio. A Methodology for the Segmentation and Identification of Music Works. In *Proceedings of the International Conference of Music Information Retrieval*, pp.271–284, 2007.

[11] R. Miotto and N. Orio. A Music Identification System based on Chroma Indexing and Statistical Modeling. In *Proceedings of the International Conference on Music Information Retrieval*, pp.301–306, 2008.

[12] R. Miotto and N. Orio. Music System for Live Performances Identification, May 2009. `http://svrims2.dei.unipd.it:8080/musicir/`.

[13] M. Müller and F. Kurth. Efficient Index-based Audio Matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, 2008.

[14] D. Turnbull, L. Barrington, D. Torres and G. Lanckriet. Five Approaches to Collecting Tags for Music. In *Proceedings of the International Conference on Music Information Retrieval*, pp.225–230, 2008.

[15] L.R. Rabiner. A Tutorial on Hidden Markov Models and selected Application. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[16] D. Turnbull, L. Barrington, D. Torres and G. Lanckriet. Semantic Annotation and Retrieval of Music and Sound Effects. *IEEE Transactions on Audio, Speech, and Language Processing*, vol.16(2), pp.467–476, 2008

[17] M. Slaney and M. Casey. Locality-Sensitive Hashing for Finding Nearest Neighbors. *IEEE Signal Processing Magazine*, vol.25(2), pp.128131 (2008)

[18] Y. Yu, S. Downie, F. Moerchen, L. Chen and K. Joe. Using Exact Locality Sensitive Mapping to Group and Detect Audio-Based Cover Songs. In *Proceedings of the 10th IEEE International Symposium on Multimedia*, pp.302-309, 2008.

Enhancing Information Management for Digital Learners

Stefanie Sieber
Media Informatics Group, University of Bamberg
Feldkirchenstrasse 21, D-96052 Bamberg, Germany
www.uni-bamberg.de/minf/
stefanie.sieber@uni-bamberg.de

Abstract

The ongoing digitalisation of the learning process not only brings many advantages for learners but also poses new challenges like the increasing amount of information that need to be faced. This paper conceptually sketches a system employing and combining different existing Information Retrieval (IR) techniques to support digital learners. The system establishes a personal learning space including all learning content by building an appropriate index of learning material. In order to store learning content and allow an additional classification, adjusted to the specific needs of a single user, two supplementary index levels are created and connected to the primary index level. Different maintenance and search facilities allow an intensive use of this individiual structure. Summarising, a personalised learning service, meeting the challenges of modern learning, is delivered by extended searching capabilities.

Keywords: personal search, personal information management

1. INTRODUCTION & MOTIVATION

The nature of learning has changed over the last decades, especially because computers and the Internet are, by now, quite naturally integrated into a typical learning environment. This technical progress and the consequential integration cause a shift from traditional learning to learning concepts located in a continuum of blended learning as defined by Jones [1]. Furthermore, not only the preparation of content itself but also the way of distributing learning material and communcation in a learning environment are in the process of being completely changed. This digitalisation of the learning process on the one hand, but also the nowadays common assumption of and awareness for lifelong learning as discussed in [2], bear new challenges, which have to be faced by learners as well as lecturers.

On the one hand, Learning Management Systems (LMS) (cf. [3]) are, in the majority of cases, employed to provide a foundation for modern learning. Whether it comes to distance learning or not, all required material, including additional information for further reading and tasks controlling the learner's success, is provided online, such that learners can easily access all desired information any time and any place. On the other hand, advanced technical possibilities also allow an integration of multimedia content into learning and teaching concepts, as for example suggested in [4]. Furthermore, regarding the concept of lifelong learning, learning is ubiquitious. Newspapers or journals are more often read on the Internet and various information is also checked on the Internet—on stationary as well as mobile devices.

Of course, this development is pleasant and softens some difficulties that usually need to be faced by learners. The general availability of information, for example, undoubtedly increased due to this technical progress. However, for this very reason, composing the big picture is more difficult, also because skimming through collected information is not satisfactorily supported by now. Moreover, the rising number of different kinds of content that needs to be considered is even more calling for an appropriate support of collecting and arranging different kinds of learning content. Besides, there is also the fact that most of the systems used so far have an institutional point of view rather

than the perspective of a single learner.
Put another way, personal and professional experiences with LMS (cf. [5]) and their functionalities, often far from being satisfactory, actually lead to proposing our system as described within this paper.

Briefly spoken, the vision of the presented approach lies in the provision of an environment, allowing a single user to administer his learning material and learning context in order to facilitate learning insights. Hence, the concept for our future system is trying to build a *personal learning space* including all objects somehow related to the learner and the qualification he is trying to attain. Exisiting IR techniques are employed and combined in order to smoothly integrate the system into the learning process of a single learner.

The rest of the paper is structured as follows: Section 2 provides a brief overview on related concepts that need to be considered when actually designing the system. Focus of attention is on the presentation of the system concept in section 3. Within this section, the three design parts, that is the main ideas behind the system components for the collection, preparation, and presentation of learning objects are described. Finally, a short outlook in section 4, providing insights into future work, concludes this paper.

2. RELATED WORK

Of course, the promotion of a personal learning space immediately suggests personal learning environments (PLEs). By general definition, a PLE is a system helping learners to control and manage their own learning: “A PLE is a single user’s e-learning system that provides access to a variety of learning resources, and that may provide access to learners and teachers who use other PLEs and/or VLEs [Virtual Learning Environments].” [6]

In order to allow a classification, Sandra Schaffert and Wolf Hilzensauer [7] identified seven crucial aspects for personal learning environments: the role of the learner as active, self-directed creator of content, personalisation, content without limitations, social involvement, ownership of learner’s data, educational & organisational culture, and technological aspects. Evaluating these criteria, most of the stated aims of the proposed system coincide, such that our system can also be classified as PLE. However, PLEs often focus on examining and actively controlling the learning process and progress, which is particularly not the main objective of our system.

Mostly, PLEs have their origins in the demand for learner-centred approaches. Forms and specific definitions of PLEs range from sets of tools, each supporting a single asset of learning, as for example examined by Dalsgaard [8], to sophisticated systems, such as those briefly described by van Harmelen [6].

Besides some current research projects like TENCompentence[1] or MATURE[2], there are, of course, several existing systems already dealing with a more general view of learning than a typical PLE. However, as far as known, there is no existing system trying to smooth the process of assembling all information related to the learning process as well as fostering new insights by providing, first, assistance for easily setting up a personal learning space, second, a sufficient search and possibilities to efficiently glance through learning material and, third, visualisations for the two previously named tasks, showing the big picture by pointing out connections of single information pieces.

Concerning the technical implementation different research areas need to be considered. First of all, the proposed system follows a user-oriented approach to IR. Therefore, developments and findings of this area, as for example described in [9], are the relevant for our system. Secondly, user interaction is an important principle of user-centred systems in general and, of course, learning environments in particular. Relying on theories of constructivism (cf. [10]) interaction is a basic principle for learning, therefore interactive IR, examined in detail by Xie [11], needs to

[1] http://www.tencompetence.org
[2] http://mature-ip.eu

be particularly considered. Thirdly, the idea of working with single informationen pieces, called learning objects, is not new. While generally working with learning objects, ideas, thoughts, problems, and solutions as for example discussed in [12] need to be taken into account. Also, storing learning objects suggests digital libaries like examined in the DELOS project[3].

Moreover, since different existing IR techniques are supposed to be combined, actual research of the specific area needs to be considered as well. Selected references are directly included in the description of section 3.

3. SYSTEM CONCEPT

The future system, trying to accomplish the vision described above, is now described in more detail. By now, the system has not been realised in full depth and detail, for this reason descriptions need to be understood as conceptual sketches. The system will be divided into three core components—a collection component, a preparation component, and a presentation component (cf. figure 1). Each component is depicted in a separate subsection.

The foundation and content of the system are *learning objects*. Basically, a learning object is any possible physical representation of information, such as locally saved files in different formats or an online web page, related to the learning process. These learning objects are assembled and traditionally stored in a database by the *collection component*. Subsequently, the *preparation component* processes the assembled objects and extracts or assigns additional information. The presentation component, in contrast, represents the interface enabling user interaction and passing results along to the user.

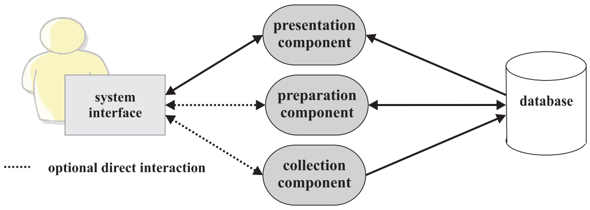


FIGURE 1: User Interaction and Interplay of System Components

Concerning the actual realization, in summary, the system is ought to adopt familiar Web2.0-paradigms such as a rich and user friendly interfaces and tagging of information assets in order to facilitate intensive usage. Additionally, the system is meant to smoothly operate in the background for most of the time by default; variations, adjustments and a more intensive usage are of course fostered for active users.

3.1. COLLECTION COMPONENT

The basic component of the components triad is the *collection component*. This component is employed to actually build the personal learning space by assembling local as well as web content by an automatic, though controlled approach in contrast to crawling the Internet in general.
Technically spoken, this component is based on a combination and extension of existing persistence and search frameworks such as Hibernate[4] and Lucene[5]. At the moment possible compositions are examined, since there are already different existing approaches and projects like Hibernate Search[6]—integrating Lucene into the Hibernate framework—or the Compass Project[7]—allowing a flexibel combination of Lucene and different Object-Relational Mapping frameworks as well as supporting Spring[8] integration.

[3] http://www.delos.info
[4] https://www.hibernate.org/
[5] http://lucene.apache.org/
[6] https://www.hibernate.org/410.html
[7] http://www.compass-project.org/
[8] http://www.springsource.org/

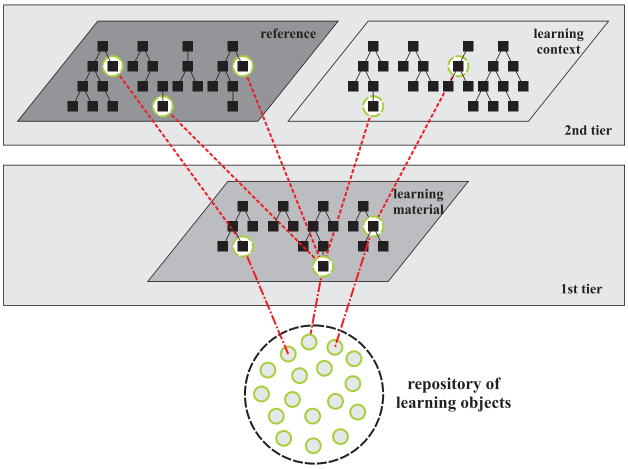


FIGURE 2: Levels of Indexing connected with Learning Objects

As core feature this component is not just extracting and processing textual information explicitly stored in *learning objects*, but generating three different index levels, building upon the database of learning objects, each of them organised in a specific hierarchy. This concept is related to the area of Hypertext IR (cf. [13]). The three levels—learning material, learning context, and reference—are loosely connected and organised in two tiers, as shown in figure 2. All different index steps can be triggered automatically for real-time indexing or manually for time-shifted adding of learning objects.

Since the technical granularity of *learning objects* is more or less arbitrary and learning objects are—at best—only structured by a manually created file structure; there is no existing structure that can be build upon. Hence, the index level *learning material*, respectively the first tier, includes all learning content but abstracts from learning objects. In this manner, a hierarchical structure linking to learning objects, however allowing file-independent connections of single learning assets, is formed. Furthermore, the supplementary index levels *learning context* and *reference* in the second tier are referencing to this particular structure. Put in other words, this level is used to abstract from physical data formats and build a well-defined hierarchy of learning material to allow precise references.
Technically, this level is similar to known indexing components. Text-based content is processed according to traditional full-text indexing mechanisms and an appropriate index is built. To successfully implement this index level as system foundation, related areas, such as XML Retrieval described by Kamps et. al [14], need to be considered.

Subsequently, the two supplementary index levels need to be populated. Therefore the collection component is trying to extract and store the *learning context* for a particular learning asset. Depending on the learner's situation, the learning material is ought to be classified as relevant, for example for a particular course or subject. Determining the learning context completely automatically is not meant to be achieved. Especially in the beginning, the detection will require additional manual effort of the user. However, preferences are supposed to be employed to structure and minimise manual input. It seems, for instance, applicable to use actual course websites for each semester in LMS or similar systems to determine the context of learning assets invoked within a certain radius from these websites.
Independent from a certain, determined learning context, learning material can be further classified by using arbitrary tags. This possibility is represented by the third index level *reference*. As mentioned, adopting the Web2.0-philosophy, the user is allowed to freely assign tags to learning material using the reference level. These tags can, if desired, be organised in one or more hierarchies in order to depict implicit structures. Here, too, manual setup and maintenance are unpreventable for a successful realisation. The fact, that this level explicitly asks for the user's input, is meant to be attenuated by offering Web2.0-like manipulation options. Additionally, the

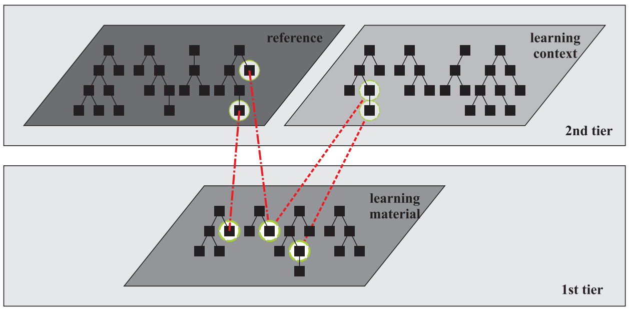


FIGURE 3: Reasoning across different Levels of Indexing

process is ought to be smoothened and supported by automatically extracting and assigning keywords like headings or captions from learning objects.

Using the three index levels organised in two tiers, as described, instead of a single index level or a database index, fosters personalization and learning insights by allowing an individual structure. As a result, the design of the collection component is closely related to faceted search and needs to consider issues of this area, as for instance examined by Henrich and Eckstein [15].

Finally, the collection component is completed by the design of an active learning mode. This mode basically activates automatic indexing. Enabling this active learning mode, all visited web pages passing a filter—validating criteria like file format or domain names—are added if a certain time threshold is exceeded. Furthermore, maintaining a white list of web pages, like course pages, that should automatically be revisited and checked for updates periodically is an absolute necessity in our opinion. These examples are just two possiblities for user defined filters, which should of course be covered by system functionalities. For local content the organisation is considerably simpler. Besides adding single files to the repository, the user should, of course, be allowed to add directories to be observed and periodically checked for updates. In this manner, it can be ensured that the learner is the active, self-directed creator of content. Regular learning actions—consuming as well as creating learning-relevant content—are employed to automatically collect the content of the personal learning space.

3.2. PREPARATION COMPONENT

Building upon the foundation created by the collection component, the *preparation component* processes and enhances the information collected and basically structured in the previous stage. The main focus is to use the three index levels for a reasoning mechanism, adopting spreading activation techniques, such as described by Crestani [16].

As depicted above, every information in the two supplementary levels of the second tier is connected to a particular node in the first tier. Hence, spreading activation models and theories can be employed to detect possible connections between single nodes of learning material that have not been obvious before. That way, learning objects not explicitly searched, however considered relevant through connections spread across the different index levels, can be recommended to the user. Figure 3 shows reasoning using connections spread across the three different indexing levels. Correlations can, for example, be used to add novel connections between learning material and tags of the third index level. The careful configuration of the reasoning component will be subject of further research and inquiries.

The preparation component, obviously, is also one possible starting point for collaboration scenarios incorporating social involvement. Especially because Web2.0 philosophy is employed and techniques like tagging are integrated, it is likely to avoid performing equal tasks by multiple users several times. The tagging of a particular learning material, for example, does not need to be computed several times but can be shared with other learners, using the same learning objects. Of course, complete ownership of learner's data needs to be retained. Elaboration of feasible implementations will also be subject to further research.

3.3. PRESENTATION COMPONENT

Finally, the *presentation component* primarily forms the investigation interface offering the service as a whole to the user. The presentation component also enables user interaction for the two previously described components by allowing manual input, collection and learning-related preparation. Direct adding, accessibility, modification, and extension of learning objects or learning material of the first tier are also provided by the interface.

For this component, it is especially important to allow a lightweight and intuitive utilisation. Principles of exploratory search, as for example described by Marchionini [17], need to be applied in order to satisfy traditional search activities like fact retrieval but in particular the continuative search activities *learn* and *investigate*.

4. CONCLUSION AND OUTLOOK

In summary, our proposed system is supposed to support the learning process of a single learner, also allowing optional collaboration with other learners if desired by employing and combining IR techniques. One of the research challenges lies in carefully selecting and combining existing techniques in order to benefit from related research areas.

Basic searching is constituted by keyword search. Of course, an advanced search, allowing to use the three index levels as filters to narrow the search down to certain criteria, also needs to be part of the described system. Among others, it is very likely that parameters like the actual learning context—determined due to the web page just visited—can be employed to find related learning objects not directly included, but assigned to the search context as well.

Moreover, different visualisation components are supposted to be embedded into the search interface for additional presentation of various aspects. Search results can be visualised by tag clouds for textual information. The collection of learning objects can—based on a particular search or not—be visualised using associations of different levels, allowing browsing the repository on a visual level.

REFERENCES

[1] Jones, N. (2006) *E-college wales, a case study of blended learning.* In Bonk, C.J. and Graham, C.R.: The handbook of blended learning: Global perspectives, local designs. 182–194. Pfeiffer.
[2] Field, J. (2006) *Lifelong learning and the new educational order* (2nd edn). Trentham.
[3] McGee, P., Carmean, C., Ali, J. (2005) *Course Management Systems for Learning: Beyond Accidental Pedagogy.* IGI Publishing.
[4] Henrich, A., Sieber, S. (2009) *Blended learning and pure e-learning concepts for information retrieval: experiences and future directions.* Information Retrieval, 12(2), 117–147.
[5] Henrich, A., Wolf, S.U., Sieber, S. (2008) *Evaluation, analysis, and future issues of a university-wide learning management system concluding a two-year initiation phase.* In Hambach, S., Martens, A., Urban, B.: e-Learning Baltics 2008: Proceedings of the 1st International eLBa Science Conference in Rostock, Germany, June 18-19, 2008. Fraunhofer IRB Verl., Stuttgart.
[6] van Harmelen, M. (2006) *Personal learning environments.* Conference on Advanced Learning Technologies, 815–816.
[7] Schaffert, S., Hilzensauer, W. (2008) *On the way towards personal learning environments: Seven crucial aspects.* eLearning Papers, (9).
[8] Dalsgaard, C. (2006) *Social software: E-learning beyond learning management systems.* European Journal of Open, Distance and E-Learning, (2).

[9] Xie, I. (2008) *User-Oriented IR Research Approaches.* In Xie, I.: Interactive Information Retrieval in Digital Environments. 1–28. IGI Global.

[10] Twomey Fosnot, C. (2005) *Constructivism. Theory, Perspectives, and Practice* (2nd edn). Teachers College Pr.

[11] Xie, I. (2008) *Interactive Information Retrieval in Digital Environments.* IGI Global.

[12] Littlejohn, A. (2003) *Reusing Online Resources: A Sustainable Approach to E-Learning.* Routledge Falmer.

[13] Agosti, M. and Smeaton, A.F. (1996) *Information Retrieval and Hypertext.* Kluwer Academic Publishers.

[14] Kamps, J., Marx, M., Rijke, M., Sigurbjörnsson, B. (2003) *Xml retrieval: what to retrieve?* In SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 409–410, ACM.

[15] Henrich, A. and Eckstein, R. (2008) *An integrated context model for the product development domain and its implications on design reuse.* In Marjanovic, D., Storga, M., Pavkovic, N. and Nojcetic, N.: 10th International design conference: DESIGN 2008, 761–768.

[16] Crestani, F. (1997) *Application of spreading activation techniques in information retrieval.* Artificial Intelligence Review, 11(6), 453–482.

[17] Marchionini, G. (2006) *Exploratory search: from finding to understanding.* Communications of the ACM, 49(4), 41–46.

Towards Personalized Advertising in Sponsored Search

Ahmad I. Zainal Abidin
Department of Computer Science, University College London
Malet Place, London WC1E 6BT, UK
a.zainalabidin@cs.ucl.ac.uk

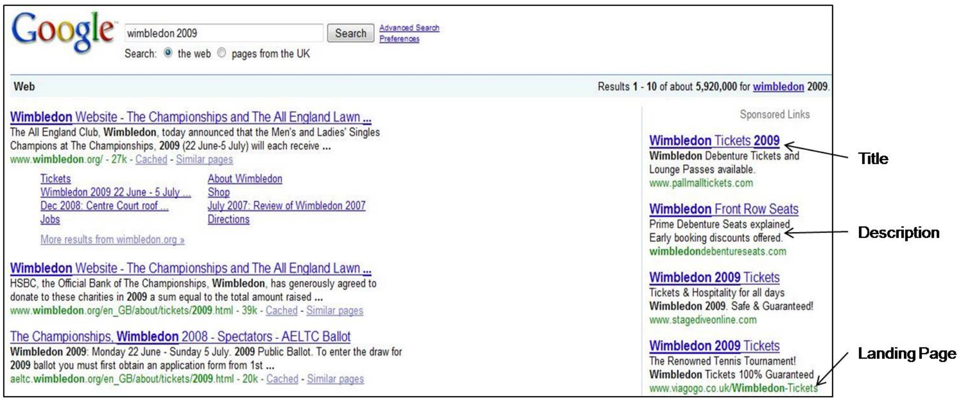
Abstract

Web advertising is one of the major sources of income for numerous search engines, news sites and non-commercial publishers. Textual ads, characterized by *Sponsored Search (SS)* and *Content Match (CM)*, make up a significant portion of Web advertising. In *SS*, with limited information about ad contents, given a query, the challenge is to place relevant ads alongside organic search results. Organic search results are ranked based on their relevance to search keyword. However, *SS* results are not necessarily ranked purely based on relevance due to various factors influencing the ads overall ranking such as bid phrase and displayed position. The displayed ads may not relate to a user's information need. In this paper, a study associating ads and users, referred to as personalized advertising is proposed. User profiles are used as external knowledge to establish the relationship between the users and the ads.

Keywords: Web Advertising, Sponsored Search, Information Retrieval, Search Engine

1. INTRODUCTION

The Web has been useful to Internet users around the world for transactional and non-transactional purposes such as information seeking, communication and online shopping. When seeking for information, a query is submitted to a search engine that later displays results relevant to the query. It is estimated that about two billion searches are performed daily by Internet users around the world [1] with various information needs. The trend of using the Web for non-transactional and transactional activities continues to increase, enabling advertisers to reach out more potential customers through Web advertising – one of the major sources of income for search engines, commercial publishers (*e.g.* news sites), and non-commercial publishers (*e.g.* blogs). Typically, the ads are distributed through either *Sponsored Search (SS)* or *Content Match (CM)* method. In *SS* (also known as *Paid Search Advertising*), textual ads are placed at the result pages for the query that a user has submitted. An example is as shown in Figure 1.



Results 1 - 10 of about 5,920,000 for wimbledon 2009.

Wimbledon Website - The Championships and The All England Lawn ...
The All England Club, Wimbledon, today announced that the Men's and Ladies' Singles Champions at The Championships, 2009 (22 June-5 July) will each receive ...
www.wimbledon.org/ - 27k - Cached - Similar pages

Tickets
Wimbledon 2009 22 June - 5 July ...
Dec 2008: Centre Court roof ...
Jobs
About Wimbledon
Shop
July 2007: Review of Wimbledon 2007
Directions
More results from wimbledon.org »

Wimbledon Website - The Championships and The All England Lawn ...
HSBC, the Official Bank of The Championships, Wimbledon, has generously agreed to donate to these charities in 2009 a sum equal to the total amount raised ...
www.wimbledon.org/en_GB/about/tickets/2009.html - 39k - Cached - Similar pages

The Championships, Wimbledon 2008 - Spectators - AELTC Ballot
Wimbledon 2009: Monday 22 June - Sunday 5 July. 2009 Public Ballot. To enter the draw for 2009 ballot you must first obtain an application form from 1st ...
aeltc.wimbledon.org/en_GB/about/tickets/2009.html - 20k - Cached - Similar pages

Sponsored Links

Wimbledon Tickets 2009
Wimbledon Debenture Tickets and Lounge Passes available.
www.pallmalltickets.com

Wimbledon Front Row Seats
Prime Debenture Seats explained Early booking discounts offered.
wimbledondebentureseats.com

Wimbledon 2009 Tickets
Tickets & Hospitality for all days Wimbledon 2009. Safe & Guaranteed!
www.stagediveonline.com

Wimbledon 2009 Tickets
The Renowned Tennis Tournament! Wimbledon Tickets 100% Guaranteed
www.viagogo.co.uk/Wimbledon-Tickets

FIGURE 1: Sponsored advertisements listed at the result page of a query

Currently, most Web advertising consists of textual ads with three elements: an ad *title*; a few lines of *description* relating to the title, promoted products or services; and a URL to the advertisers site called the *landing page*. Major search engines such as Google and Yahoo! generate income through textual ads placed at their search results. Basically, they act as search engines that also provide ad networking services. In the United States, search advertising shared about 45% of the total $23.4 billion generated by Web advertising for the full year of 2008 with overall Web advertising revenue for that year increased nearly 11% over 2007. The remaining 55% of the total $23.4 billion was generated by a combination of display-related advertising (33%), classified advertising (14%) and other advertising formats (8%) [2]. Search advertising is therefore in an advantageous position for advertisers to reach out customers and for search engines to generate advertising revenues. Internet users who initially use the engines to seek information are the targeted group to interact with the advertisers' ad campaigns. The format of *SS* advertising is less annoying compared to unsolicited commercial emails.

The second method of textual ads in Web advertising is *CM* (or *Contextual Advertising*) which places ads within the content of a Web page. In *CM*, an ad network is involved with the goals of optimizing relevant ads and revenues shared between the Web page owner (*e.g.* blogger) and the ad network. However, *CM* is not the focus of this paper.

Advertisers bid for keywords, known as *bid phrase*, to be associated with their ads. The position where the ads are being displayed is also a factor in determining how much the advertisers are being charged. The three pricing models for textual ads are *pay-per-click (PPC)*, *pay-per-impression (PPI)*, and *pay-per-action (PPA*). In *PPC* pricing model, advertisers will be charged a certain amount for any clicks on the URLs provided in their ads. For *PPI*, advertisers will be charged every time their ads are displayed, and for *PPA*, advertisers only pay if users visit the advertisers Web page and later perform transactions.

Although the displayed ads relate to the query that a user has submitted, the ads do not necessarily relate to the user as a potential customer. For instance, suppose that users *A* and *B* type *"wimbledon 2009"* as a query, *A* may be interested in one of the ads shown in Figure 1, but *B* may not. The reason would be that *B* likes travelling in a group so he or she is looking for a tour package not just a ticket. Apparently *A* and *B* have different profiles. Motivated by this problem, a study on personalized advertising in *SS* is worth carried out.

This paper focuses more on *SS* advertising and aims at providing an overview of techniques relevant to *SS*. The organization of the paper is as follows. Section 2 describes an overview of *SS* advertising, section 3 related works to *SS* advertising, section 4 future challenges, and section 5 concludes the paper.

2. SPONSORED SEARCH ADVERTISING

Sponsored search advertising mainly concerns about marketing products and services thought to be of interest to a user. Although the user's interests are subjective, a query the user submits to the search engine represents his or her information need. The difference between *SS* and brand marketing is that the latter promotes specific brands while *SS* promotes products and services directly based on prior knowledge, *i.e.* query keywords.

Typically, *SS* involves three entities namely *advertiser*, *user*, and *search engine* which each has a different goal:

- The *advertiser* provides ads that each consists of a *title*, *description*, and *landing page*. The goal is to promote products and services within a specific campaign period or temporal constraint so that potential customers are reached out as many as possible. For example, Sponsored links in Figure 1 revolves around Wimbledon 2009, in which one of the advertisers sells Wimbledon tickets. So, the ad theme is relevant to the user's query.
- The *user* views ads, and if he or she becomes interested, he or she will visit the advertiser's web page. The user has an information need and if the ads are relevant, the probability that the user will interact with the ads is higher compared with if the ads are not relevant. A study

in [3] confirms that the relevance of ads with user interest is important. In [4], it is evident that the ad and its context have a significant effect even when a conscious response such as click is not present.

- The *search engine* selects ads based on the user's query and places them on the result page. The goal is to select relevant ads so that users might be interested in the ads. As the search engine earns advertising revenue, maximizing click-through rate is therefore desirable. In this case, the pricing model is assumed to be *PPC*, in which when a user clicks an ad, the advertiser is being charged.

User profiles are important so that ad campaigns will be more personalized across different users. Given different users submitting the same search keyword relatively at the same time, all users should view different ads depending on their profiles. Basically, an advertiser interacts with a search engine ad networking service to add or update its ads. The engine stores the ads in its storage. A user interacts with the search engine that records the user's Web browsing and search activities. These activities are useful in formulating the user's profile that is used as an additional knowledge to relate different users with different targeted ads.

The click revenue R that a search engine can estimate for a given query q and a set of ads $A = \{a_1, a_2, ..., a_n\}$ is expressed as

$$R = \sum_{i=1}^{n} P(click|q, a_i).Price(a_i, i) \tag{1}$$

where $P(click|q, a_i)$ is the click probability given the query q, the set of ads A, the total number of ads displayed n and $Price(a_i, i)$ is the ad a_i click price at position i.

The main challenge with *SS* advertising is that a query is short, typically 2-5 terms, and ad contents are limited. Initially, when a user submits a query, he or she anticipates the best search results not the best ads associated with his or her query terms. Due to the brevity of the query terms and the ad contents, the best scenario would be an *exact match* between the ad bid phrases and the query terms.

One of the possible approaches to rank query q with regard to the ad a_i is by computing their cosine similarity:

$$sim(q, a_i) = \frac{\vec{q} \ . \ \vec{a_i}}{\left|\vec{q}\right| \times \left|\vec{a_i}\right|} = \frac{\sum_{i=1}^{n} w_{iq}.w_{ij}}{\sqrt{\sum_{i=1}^{n} {w_{iq}}^2}\sqrt{\sum_{i=1}^{n} {w_{ij}}^2}} \tag{2}$$

Although many bid phrases may be associated with an ad, it is impractical for advertisers to provide an exhaustive list of bid phrases for their ad campaigns. Therefore, a *broad match* is an alternative to the *exact match* where a query is related to but does not have to match the phrases. In *broad match* advertising, the criterion is relaxed by not relying solely on the phrases. For example, the ad title, description, and URL for the landing page are typically used to be matched against the query. Additionally, information related to the user or the advertiser may also be considered.

3. RELATED WORK

Broder et al. [5] use a global score threshold in their work to predict whether the entire set of ads is relevant to be displayed. For any query, ads with scores higher than the set threshold are perceived to be more relevant than those with lower scores and ads with scores lower than the threshold should not be returned. So, different thresholds will result in different suggested number of ads to be returned. Clearly, there is a trade-off between the effectiveness and the coverage of ad results. The drawback of this method is that it is difficult to set a threshold that leads to a set of optimal ads because the result coverage also depends on the ad corpus. The follow up work uses machine learning technique to classify whether an entire set of ads is relevant [5]. The input

TABLE 1: Categorization of methods for ad selection

Approach	Representative Works
Machine learning	Predicting whether the entire set of ads is relevant to be displayed [5] Using click data for training and evaluation, determining which framework is more suitable, and determining useful features for existing models [9]
Threshold	Global threshold to determine whether to return ads [8]
Query expansion, augmentation and substitution	Rewriting of tail queries [6] Using knowledge in search result to create augmented query [8] Optimizing ad relevance and revenue using query substitution [11]
Click-through and click feedback	Estimating the click-through rate [12]

is a query and a set of ads, while the output is either "yes" if the entire set should be displayed or "no" otherwise. In their classification model, they use Support Vector Machine.

In *SS*, query rewriting, done mostly offline using various sources of external knowledge, is a common technique used to perform *broad match*. Repeating queries from the head to torso of the query volume will benefit from this approach. Online query rewriting for *SS* was proposed in [6]. In their work, they use pre-processed queries to develop inverted index of the expanded query vectors and run incoming queries against it to enrich the original tail query representation. In their work to rewrite queries for *SS*, Jones et al. [7] used search engine query logs to gather user session information and later produce alternative queries for ad selection. Candidate substitutions are first generated by examining pairs of successive queries that user issue in the same session. Subsequently, the candidates are examined to find common transformations. A machine learning ranking is used to determine the most relevant rewrites that match against ads. In [8], the authors use the search results (*top-n* pages) so that an additional knowledge about the query submitted by the user can be gathered. The additional knowledge is then used to augment the query. The augmented query is later evaluated against the ad corpus to retrieve relevant ads. In [9], the authors use machine learning approach to predict the likelihood of an ad is to be clicked. Knowing which ads likely to be clicked is very important to both the advertisers and the search engines.

Given a set of relevant documents and a set of irrelevant documents, the work on query expansion by Rocchio [10] shows how the best query is crafted. With regard to *SS*, documents refer to ads. Rocchio's expanded query consists of a linear combination of the original query, and the vectors for relevant and irrelevant documents. The vectors for relevant documents are added to the original query, while the vectors for irrelevant documents are subtracted from the original query. Radlinski et al. in [11] propose a query substitution approach to optimize relevance and revenue in *SS*. They analyze frequent queries in pre-processing phase that constructs lookup table which will map queries into bid phrase. User original queries are transformed to substituted queries that have higher matching chances with bid phrase. Richardson, Dominowska and Ragno in [12] propose a method to estimate the click-through rate for new ads. They develop a model using logistic regression. Their dataset consists of information on landing page, bid term, title, body, display URL, clicks, and views for each ad. The dataset is used to train and test their regression model.

Recent research works on *SS* advertising are classified in Table 1.

4. FUTURE CHALLENGES IN SPONSORED SEARCH

The problem in Web advertising will remain chiefly on selecting relevant ads for a given query especially the ads volume is expected to increase. However, the research and applications in Web advertising are dominated by search engine companies due to advertisement datasets and ads bidding networks are available to these companies. Search engines generate the ads relevance and ranking by their proprietary algorithms which lack public evaluations or involvement of researchers in academia.

When a user submits a query, the expectation is to have the best search results based on their relevance to the query. *SS* results are influenced by various factors: the relevance between the user's query and the advertiser's bid phrase, the bid amount for the phrase, and the position chosen to display their ads. Each of these factors has its own weight in determining the ads overall ranking. Therefore, studying the trade-off between ads relevance and advertising revenue is essential. An ad positioned in the top list may be due to the dominant weight assigned to its bid phrase. The click-through data are worth analyzed to study the number of viewed ads that are later clicked regardless of their displayed positions.

While research works on query expansion and query substitution have been done to augment or substitute original query to improve ads relevance, the displayed ads may not necessarily be relevant to users' interests. Therefore, given the same query, two different users with different Web browsing and search interests should be displayed with different ads. Therefore, learning consumer behaviors that represent interests will be important in personalized advertising. Click-through logs that have the data on users' web browsing and search interests can be used to model behavioral profiles. However, monitoring users' web access raises a privacy concern. If their Web access activities are being monitored for marketing purposes, users should be made aware of the monitoring usage. Perhaps, users should be rewarded for allowing their web access activities to be monitored. A study on the technical, legal and ethical implications of behavioral profiling needs to be addressed.

Most researches in Web advertising focus on issues related more to search engines. To the best of my knowledge, there has not been much research on advertiser-oriented Web advertising in the literatures. Generally, advertisers have the budgets to run their campaigns. It is important to optimize advertisers' satisfactions and to minimize their advertising costs. The metrics to measure the advertisers' satisfaction need to be developed.

5. CONCLUSION

This paper provides an overview of Web textual advertising especially on sponsored search, the recent research works in sponsored search and its potential future directions. The author views that personalized advertising in sponsored search is essential because different users will view ads relevant to their interests and information needs. It is known that an advertising revenue is a driving force in sponsored search. So, the displayed ads in sponsored search may not necessarily be the most relevant ads to a user's query or interest. Therefore, improving ads relevance that relate to users as potential customers may be more rewarding to the advertisers than displaying ads that no users relate to.

REFERENCES

[1] 61 billion searches conducted worldwide in August. October 2007. Available from http://www.comscore.com/press/release.asp?press=1802

[2] IAB Internet advertising revenue report: 2008 full-year results, March 2009. Available from http://www.iab.net/media/file/IAB_PwC_2008_full_year.pdf

[3] C. Wang, P. Zhang, R. Choi, and M. D'Eredita. Understanding consumers attitude toward advertising. In *8th Americas Conference on Information Systems*, 2002

[4] C. Y. Yoo. Preattentive processing of Web advertising. *PhD Thesis*, University of Texas at Austin, 2005.

[5] A. Broder, M. Ciaramita, M.Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras, To swing or not to swing: learning when (not) to advertise. In *CIKM08*, page 1003-1012, 2008

[6] A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, D. Metzler, L. Riedel, and J. Yuan. Online expansion of rare queries for sponsored search. In *WWW09*, 2009

[7] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW06*, page 387-396, 2006

[8] A. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. Search advertising using web relevance feedback. In *CIKM08*. Page 1013-1022, 2008

[9] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In WWW08, page 227-236, 2008

[10] J. J. Rocchio. (1971). *Relevance feedback in information retrieval.* Prentice Hall: Englewood Cliffs, New Jersey

[11] F. Radlinski, A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel. Optimizing relevance and revenue in ad search: a query substitution approach. In *SIGIR08*, page 403-410, 2008

[12] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW07*, page 521-529, 2007

A Multi-disciplinary Approach to Interactive Information Retrieval upon Semi-structured Data Sets

Corrado Boscarino
Centrum Wiskunde & Informatica (CWI), Science Park 123,
1098 XG Amsterdam, The Netherlands
corrado@cwi.nl

Abstract

The so called logic and probabilistic views on IR can be reconciled by a unifying framework for IIR. I present a proposal for a PhD research according to a multi-disciplinary perspective and I discuss some of its consequences for IR as a discipline.

Keywords: PhD proposal, Interactive Information Retrieval, dynamic logics

1. INTRODUCTION

Richard Feynman [7] once said that the existence of barriers between disciplines is one of the main obstacles to the progress of science. Conceptual walls are built even inside one single discipline, like between Database-systems (DB) and Information Retrieval (IR) [4] in the burbling pot termed as a whole 'Computer Science', only to find out that they are doing more harm than good to the scientific cause. In the meantime, while the majority of practitioners still believes in the value of a kite-mark for permissible approaches, dwelling in between this received network of disciplines and specialisations is often regarded as an extracurricular activity, which you should not engage with at the expense of the institution that appointed you. In the most fortunate cases it is considered to be a matter of chiefly academic interest, which usually means that only tenured professors may safely pursue that path.

Things start to change when an external event, often the emergence of a technological artefact, confronts us with a paradigm's precinct. For example, the development of nano technologies [14] does not merely improve on the existing engineering practice, but introduces radically different procedures for the manipulation of materials at the atomic scale; the fraction of surface atoms becomes comparable with that of the bulk, the thickness of a layer of material approaches the wavelength of the electronic functions. At this critical scale, material scientists are concerned with quantum effects, chemistry and electronic engineering become deeply intertwined and biology sees the production line as a viable perspective. Practitioners then become aware that landmarks in their own discipline are rough approximations in another discipline, which stay valid until a limit is reached, whether conceptual or technological.

When a larger amount of computing power and advanced data processing applications are accessible by a greater part of the population, as is now the case in Western societies, information becomes a commodity: extraction, processing and distribution of bulk information are not simply extensions in their scope of techniques for accessing structured information repositories, but they are different in kind. Traditional disciplinary boundaries disappear at the scale of indeterminacy reached by modern applications: the amount of disorder in the structure of information repositories and information needs plays in IR the same role as a scaled parameter in physics. Whereas physical laws depend on the scale of parameters such as size, speed or force, methods in IR depend on the amount and kind of structure that informs both the information sources and the queries.

Semi-structured data sets are not less structured deterministic repositories that can be processed by adapting the tools for dealing with structured repositories; their lack of structure blurs the distinctions between IR, DB, and also between these sub-disciplines of computer science, and those fields of philosophy, sociology, anthropology, physics and logic that are concerned with information, its meaning and its use by human agents.

The next section shows that when, information appears to be less and less ordered, and therefore more complex, a successful approach to description, processing and retrieval of information cannot be exhausted within a single discipline. In particular I consider three realms: i) probabilistic models, which realise mappings under the rules of probability theory from the form of the information repository to the form of an information need, ii) semantics, which is considered to either partake only the truth conditions of sentences in a language or more widely to refer to any formal representation of the meaning carried by an informational structure, and iii) user interaction, which may be provisionally termed as any flow of information between a human user and an IR system.

A multitude of disciplines provide valuable insights on all these issues, however applying to IR different conceptual frameworks developed elsewhere would reaffirm the status of an alleged disciplinary essence. In section 3 I shall then introduce my proposal as a demand for disciplines to rethink their theoretical foundations. IR challenges the existing paradigm and we cannot simply resolve to apply different probabilistic models to the problem of evaluating the relevance of a set of documents with respect to a certain query, but we are prompted to address important issues of probability theory through IR. We cannot simply apply theories of meaning to IR, but IR concurs to the development of new theories of meaning. The research proposal that I sketch here has the twofold aim i) to provide a common framework where current approaches to IR, to the description of information flows and to user interaction can be discussed and ii) to provide a design framework where novel applications can be developed, taking advantage of the synergies between different disciplines and improving on the existing theories and practices. While I cannot possibly be exhaustive in the range of applications and on the potentiality of this framework, I will simply present one motivating example of theories, which are known to have been successful in accounting for information exchange related issues, but which are also difficult to reconcile, mainly because they have been developed within the scope of different disciplines.

2. A CURRENT ISSUE IN IIR: SEMANTICS AND PROBABILITY

A large number of insights about the three parts that I regard IIR to be mainly made of are already provided by a variety of disciplines, albeit they are loosely or completely unrelated to each other. In this section I present an issue, which preliminary investigations show that can be tackled with the framework that I aim to. It is related to current debates in the broader field of information access and serves as a starting point for my research.

This exemplary case concerns the long-lasting question whether IR should be concerned with the meaning of information, whatever notion we want to append to this concept, and be about the development of tools designed to ‘understand’ the content of both the documents and the queries and to find suitable mappings between the two. Alternatively, IR can deal only with the outer form of the data, leaving interpretation to the user, provided that enough data are available to make probabilistically responsible statements about relations, further unspecified in their interpretations, between the statistical parameters of the documents and those of the queries. Even at the time of writing, this debate heats the feelings of those involved in information access research [11]; the question whether we should renounce altogether to attempts to model human reasoning only because it is clearly too complicated to be grasped by mathematically elegant expressions is still unsettled.

Without doing too much harm to the multiformity of the different positions, we can cluster a first view on the debate about the relation between semantics and probability around Fuhr’s survey of probabilistic methods for IR [9], essentially based on the application of probabilistic reasoning such as Bayesian methods and on the quantification of relevance as the lumped parameter of user satisfaction. To the same class of probabilistic methods belongs the language modelling approach [12], which seeks to determine the unique model of the document that generated a query, modulating natural language processing techniques to the task of IR. There have been even attempts to show that the two are rank equivalent [5, pp. 1-10]. Although I do not want to subscribe to this thesis, which has been proved to be in itself problematic [16], the two probabilistic approaches are equivalent in how they put themselves in relation to semantics: since there are a limited number of possibilities to compose the syntactic and lexical units and still produce the same semantic structure, an IR application may simply aim to match those structures, leaving the user to discriminate between the residual subtleties.

Conceptually different are approaches such as that of Nie [15], and before that of Van Rijsbergen [20], who seeks to add semantic awareness to classical probabilistic models. This is at the same time a bridge to IIR in that it enables the employment of different semantic notions for IR purposes, included those that regard meaning to be tightly coupled to human experience. Although this logical approach to IR forms the main anchor to my research, the proposed techniques to convert the output of logic processes to probabilistic statements are still lacking a thorough framework going beyond an *ad hoc* solution for the case of logical implication. The two classes of approaches are seen as competitive only because, as Wong shows [22], IR models are probabilistic implementations of processes of logical inference; when logics are thought of as being mainly concerned with inference we find the logic and the probabilistic view on IR on two different sides if we focus on the implementational layer and to be equivalent at the conceptual level; in both cases it is difficult to design applications where the two views coexist and complement each other.

Relevance may still hold as lumped parameter, but its relation with probabilistic relevance is not easy to determine. The consequence is that relevance can only be characterised in relation to a particular IR system [2], which is conceptually unsatisfactory as we would expect it to be related to the context, the particular query, the previous information gathered by the user, but less to the particular technique that one uses to access information. A rough simplification may lead to affirm that those different approaches to IR are just different estimations of relevance, albeit without a clear understanding of how this parameter is bound to human potentiality. The picture that we get is that of a claim that a certain mathematical expression models human satisfaction without an understanding of what this satisfaction is about.

In order to solve this bottleneck of the design *within* the IR community, without addressing, for example, how anthropologists generalise observations of a local, small scale community, to general cultural theories, very advanced evaluation protocols [21] have been developed. This transfer of theoretical models from design to evaluation does not prevent the closure of the discipline and the elaboration of notions, like that of epistemic uncertainty in [23], specifically targeted to the problem of IR evaluation, although they could be extended to a wider scope. Bringing the human user into the design of IR applications and yet limiting the scope of the theoretical analysis to the boundaries of IR as a discipline also leads to unsatisfactory results. In [11] Fuhr presents a theoretical model for IIR in which he is forced to subscribe a rather strict set of assumptions only in order to let the model be compatible with classic probabilistic IR. In particular concepts such as information need or relevance are left unquestioned.

The few notable exceptions of cross-fertilisation of computer science with theoretical insights from other disciplines remained halfway towards the approach I suggest: they allowed the design of technological artefacts to be inspired by other fields without actively contributing to the other discipline. This is the case of [6], which discuss Merlau-Ponty's phenomenology or Wittgenstein's theory of meaning, both well known and perhaps also slightly outdated, without contributing to the productive field of anthropology of the senses or to that of pragmatics in philosophy of language. IIR should both develop its own theories and communicate them so to trigger advances in the humanities, affording novel ways to do anthropology or philosophy *through* IIR; this is an important benchmark to assess the quality of the solutions developed in such an multi-disciplinary research line.

Before sketching in the next section how a framework that pursues this path further may look like, I shall conclude the present discussion with a methodological remark.[1] It appears that the preconditions to design a multi-disciplinary framework for IIR relies heavily on the expected results of the framework itself; if it does not fall pray of circularity, at least this approach can be facilitated by some bootstrapping procedure. The specialisation of modern research does not allow even a large team to encompass all the different skills that may be relevant to the task. One possible solution to this problem could be to introduce a meta-framework, like that suggested in [1], that provides the key concepts to bridge disciplines. This is, however, still an open issue, which is faced with some regularity in the design of many application in information science. Without this bootstrapping procedure at our disposal, we are then forced to manually select the concepts that, to the best of our knowledge, have been at the core of different disciplines and to use them during the development of the framework.

3. RESEARCH DIRECTION AND CHALLENGES

A probabilistic model of IR can be thought of as a functional block that provides a mathematical formalisation of reasoning under uncertainty and that can be interfaced to other blocks, inducing a hierarchy of different models, each with its specific accessory functionality. A statistical components block attached to the probabilistic model specifies how probabilities should be calculated, one or more logical blocks specify our knowledge about the data sets, the context, the user and so on; logical blocks enter the model through its priors [3], which allow the incorporation of issues that cannot be resolved at the probabilistic level, but which nevertheless may be crucial to the success of the retrieval task. This design scheme overcomes the distinction between the logical and the probabilistic view on IR.

Not yet having an automated method to import other models into the hierarchy, we are then forced to posit a theoretical background, discuss it outside that framework, which still must be designed, and let artefacts based on the framework interact with the probabilistic model; this approach resembles that followed in [13], where, however, little space has been reserved for discussion, such as why the Cognitive framework has been chosen and how it relates to other theories. I also put an additional constraint on the choice of the theoretical approach, that should admit a mathematical representation: either the framework itself must already be expressed in mathematical terms or it must be possible to couch at least its main components in a mathematical language.

A theoretical framework for IIR upon semi-structured data sets should account for the human driven process through which a set of determinately true or determinately false statements can be associated to data sets even when they lack a complete formal structure. Adding the quality of multi-disciplinarity requires also to assess how different theories explain key concepts like meaning, uncertainty or action. Finally, a probabilistic implementation will then assign probabilities to interpretations, given the data set, accessible through a statistical components block, and the procedural and structural information provided through the priors. This section explains, by means of two examples, how different insights on how meaning arises from unstructuredness and interaction, can be integrated towards a common theoretical framework for IIR. For each possible theoretical solution I will provide some clues on how a modular design of novel applications can be implemented in practice.[2]

The first example is an extension of classical logic, which is mainly concerned with assertions and with formalising the task of making an inference, that is making explicit some informations already present within a knowledge base, towards a dynamic logic for IIR. Our task requires a notion of meaning as the product of a dynamic process of interaction between the user, whose capacities far exceed that of drawing inferences, and the external world. User-system interaction and IR is then primary with respect to any crystallised information and knowledge representation structure: meaning does not arise at the system or at the subject in isolation, but at the relation the people

[1] I would like to thank the anonymous reviewer, who pointed out this fundamental issue.

[2] The reader without a background in logic and philosophy may want to skip the technical details.

create with other people, objects and ideas through technology. Multi-agent communication, while it is not commonly regarded in terms of inference

> fall[s] squarely within the scope of modern logic, viewed as a general account of information flow. To emphasise the point, asking a question and giving an answer is just as 'logical' as drawing a conclusion! [19]

The input provided to the probabilistic model through its priors is a characterisation of the dynamics of meaning within a retrieval session. Empirical or theoretical arguments may lead to the identification of different informational events that affect both the context and a user's epistemic state, provided, as I already pointed out, that we are able to produce a mathematical formalisation. Let us suppose that we want to enhance the probabilistic model with a formalisation of the event $!\phi$ of asserting that ϕ in such way that every user, who has access to the system knows that ϕ and she is able to update her epistemic state with that information. Obviously, adjusting the priors with this information leads to a modification of the posterior distributions, for that retrieving the information that ϕ is not relevant any more, the probability of a user asking ϕ will be low, and so on.

A logic block that provides this functionality could be based on a public announcement logic PAL [18], eventually enhanced with common knowledge towards a logic PAL-C. The theory has a mathematical form, hence it is implementable, because the language has a model-theoretic semantics defined onto a model $\mathscr{M} = \langle W, R, V \rangle$, where W is a set of possible worlds, R an accessibility relation from the set of users to the powerset of W^2 and a valuation V fixes the interpretation of variables. As consequence of the announcement ϕ the probabilistic model is updated through the recalculation of the priors that follows each update $\mathscr{M}|\phi$ of the logic model by ϕ. The interpretation of the proposition ϕ can vary in the different blocks, which may be attached to the probabilistic model and, in case of a PAL-block, will be $[\![\phi]\!] = \{v \in W | \mathscr{M}, v \models \phi\}$, that is after the announcement that ϕ, which is supposed to be veridical, only worlds where ϕ is the case are further considered in the calculation of the priors.

Other blocks can be designed, which formalise the acquisition of various structural or procedural informations, thereby determining additional constraints to the possible values of the priors. A functional block that formalise our knowledge about how a IR session proceeds, could be based on a logic of interrogation LoI [10], which admits also a model-theoretic interpretation where an interrogative $?\phi$ partitions the set of worlds into subsets with the same truth value for ϕ and an assertion $!\phi$ selects that partition where the ϕ is the case. Blocks based on temporal logics may formalise the temporal dimension of querying the system and many more can be designed.

The second example, which concludes this section, is an application of anthropological philosophy to the task of characterising a human user, who engages in IR and the way she interprets documents or queries, which do not have a well specified structure. This approach is mainly rooted in Badiou's theory of the subject, but admits, as Fraser shows in [8], also alternative descriptions in terms of intuitionistic logic and Kripke models. Both the meaning and the probabilistic models of complex resources are not expressible in closed form, hence the final epistemic state of the subject and the models are infinite. Also in this case we must seek to determine a formalisation that is expressed in a mathematical language in order to guarantee the feasibility of the calculation of the priors.

The suggested representation is in the form of a generic extension, a finite model extended by a generic set. The concrete documents and queries, being always finite are thought of as approximations by forcing conditions, through which statements about the generic extension are reduced to statements about the finite model. An application of Badiou's idea that the subject is mathematical, connecting an event to elements of the generic set leads to overcome the difficulties in defining the meaning of very complicated resources and the probabilistic models that definitely mark them as relevant with respect to a certain query. Following the exposition in [8], the domain of the subject is described by a set of conditions ©, which is at the same time an element and a subset of the fundamental situation S and $\pi_1, \ldots, \pi_n$ are the sets that define the conditions of ©. The goals of a subjective procedure is a generic truth in the form of a 'correct subset', which is governed by the rule Rd_1, which states that if $\pi_i \in \delta$ then $\forall \pi_j \subseteq \pi_i \Rightarrow \pi_j \in \delta$ and by the rule Rd_2, which states that, between any two conditions π_i, π_j that belong to the correct subset holds that either $\pi_i \subseteq \pi_j$ or $\pi_j \subseteq \pi_i$; a sufficient condition to ensure that the latter *compatibility* relation holds is that $\forall \pi_{i,j} \in \delta,\ \exists \pi_k \in \delta:\ \pi_{i,j} \in \pi_k$.

The next step towards a mapping of the procedural information about the interpreting subject, consists into defining a spread Δ of correct subsets over © and to add indexes to the set δ in order to indicate the position of the subsets onto a partial order. Defining a spread over © amounts to fix a function that implements the two laws $Rd_{1,2}$ that has $\wp(\copyright)$ as domain and whose range is $\{0, 1\}$. We may consider a potentially generic sequence in Δ, which can arise from a free choice sequence that never becomes expressible by an algorithm. As pointed out by Fraser, we should pay much care to regulate the model in such a way that a certain balance can be reached between the need of keeping the subject free on the one hand and to avoid a restricted law-like sequence on the other hand. This requirement can be satisfied by defining a set ♀such that for every law-like sequence λ, $\lambda(n) = \text{♀}(n) \Rightarrow \exists m:\ \lambda(m) \neq \text{♀}(m)$; while it does not allow to definitely indicate at an intermediate step of the algorithm, whether the procedure is generic, it keeps at any point the free choice sequence at a distance from the law. The problem in implementing this block is that any truth procedure is inherently anticipatory and there is no empirical gathering technique that can possibly encompass the entire sequence. At this point, Fraser introduces *forcing* as a means to save the intermediate results of the subjective enquiries. The function of the block is to gradually make sense, by proposing suitable hypothesis and gathering the necessary empirical evidence, of the initially unintelligible terms of what Badiou terms the subject-language, through the forcing relation. Since I present these results in the form of a direction for future research, it is far outside the scope of this introduction to present a complete axiomatisation of the logic of the subject. It suffices here to say that the forcing relation bears a similarity with entailment within an intuitionistic setting; in particular, a set π may force a formula $\neg\neg\phi$ of the language, and yet not force ϕ or it is possible for π to force $\neg\phi$ as long as no other condition of the same generic sequence forces ϕ. What it is also not possible to discuss here is the relation, explained in [17], between intuitionistic, and hence subjective, logics and modal logic; this resemblance has already been observed by Badiou in case of negation that may hold until another construction shows that the positive formula is the case.

4. CONCLUSIONS

After making an appeal for multi-disciplinary research in IIR and for the reconciliation of the probabilistical and the logical view on IR, I introduced a framework where functional blocks, obtained by the mathematical formalisation of different theories, are interfaced with a probabilistic model. I sketched two possible designs: one, which applies dynamic logics, yields the formalisation of public announcements and constraints on the IR system, the second one, formalises the multiplicity of the subject by means of two mathematical structures, the generic set and the forcing constructions. The arguments that I presented in favour of the need for a framework to both discuss different views and to develop novel applications in IIR, seen in its widest acceptation as a technological artefact designed to aid humans in collecting information according to certain criteria, are far from comprehensive of all the perspectives under which we can look at this subject. Nevertheless, I believe to have provided enough support to the claim that the approach that I propose accounts for the complexity of IIR and that only by bundling the strengths of logic, probability theory and philosophically sound theories on human agency can we attain a deep understanding of this subject.

REFERENCES

[1] U. C. Beresi, M. Baillie, and I. Ruthven, “Towards the evaluation of literature based discovery,” in *Proceedings of the workshop on Novel Evaluation Methodologies (at ECIR 2008)*, M. Sanderson, M. Braschler, N. Ferro, and J. Gonzalo, Eds., 2008.

[2] P. Borlund, “The concept of relevance in IR,” *Journal of the American Society for Information Science and Technology*, vol. 54, no. 10, pp. 913–925, May 2003. [Online]. Available: http://dx.doi.org/10.1002/asi.10286

[3] C. Boscarino and A. P. de Vries, “Prior information and the determination of event spaces in probabilistic information retrieval models,” in *Proceedings of the 3rd International Conference on Theory of Information Retrieval (ICTIR 09) - Studies in Theory of Information Retrieval*, 2009, to be published.

[4] S. Chaudhuri, R. Ramakrishnan, and G. Weikum, “Integrating DB and IR technologies: What is the sound of one hand clapping?” in *Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR 05)*, M. Stonebraker, G. Weikum, and D. DeWitt, Eds., VLDB Foundation, ACM SIGMOD. Asilomar, CA, USA: VLDB, 2005, pp. 1–12.

[5] B. W. Croft and J. Lafferty, *Language Modeling for Information Retrieval (The Information Retrieval Series)*. Springer, December 1999.

[6] P. Dourish, *Where the Action Is : The Foundations of Embodied Interaction (Bradford Books)*. The MIT Press, September 2004.

[7] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics including Feynman’s Tips on Physics: The Definitive and Extended Edition*. Addison Wesley, August 2005.

[8] Z. Fraser, “The law of the subject: Alain Badiou, Luitzen Brouwer and the Kripkean analyses of forcing and the Heyting calculus,” in *The Praxis of Alain Badiou*, P. Ashton, A. Bartlett, and J. Clemens, Eds. Elsevier, 2006. [Online]. Available: http://www.re-press.org/content/view/21/38/

[9] N. Fuhr, “Probabilistic models in information retrieval,” *The Computer Journal*, vol. 35, no. 3, pp. 243–255, 1992. [Online]. Available: citeseer.ist.psu.edu/fuhr92probabilistic.html

[10] J. Groenendijk, “The logic of interrogation: classical version,” in *Proceedings of the Ninth Conference on Semantics and Linguistics Theory (SALT-9)*. CLC Publications, 1999.

[11] A. Halevy, P. Norvig, and F. Pereira, “The unreasonable effectiveness of data,” *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[12] D. Hiemstra, “Using language models for information retrieval,” Ph.D. dissertation, University of Twente, Enschede, January 2001.

[13] P. Ingwersen and K. Järvelin, *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[14] Madia and J. William, “Building the future an atom at a time: Realizing Feynman’s vision,” *Metallurgical and Materials Transactions B*, vol. 37, no. 5, pp. 683–696, October 2006.

[15] J.-Y. Nie, “Towards a probabilistic modal logic for semantic-based information retrieval,” in *SIGIR ’92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 1992, pp. 140–151. [Online]. Available: http://dx.doi.org/10.1145/133160.133188

[16] S. Robertson, “On event spaces and probabilistic models in information retrieval,” *Inf. Retr.*, vol. 8, no. 2, pp. 319–329, 2005. [Online]. Available: http://dx.doi.org/10.1007/s10791-005-5665-9

[17] J. van Benthem, “The information in intuitionist logic,” to appear in Synthese special issue on Philosophy of Information edited by Luciano Floridi & Sebastian Sequoiah-Grayson (Oxford). [Online]. Available: http://staff.science.uva.nl/johan/

[18] J. van Benthem, J. van Eijck, and B. Kooi, “Logics of communication and change,” *Inf. Comput.*, vol. 204, no. 11, pp. 1620–1662, 2006.

[19] J. v. van Benthem, “Logic and the flow of information,” University of Amsterdam, Tech. Rep., 1991.

[20] C. J. van Rijsbergen, “A new theoretical framework for information retrieval,” *SIGIR Forum*, vol. 21, no. 1-2, pp. 23–29, 1987.

[21] E. Voorhees, “The philosophy of information retrieval evaluation,” in *In Proceedings of the The Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*. Springer-Verlag, 2001, pp. 355–370.

[22] S. K. M. Wong and Y. Y. Yao, “On modeling information retrieval with probabilistic inference,” *ACM Trans. Inf. Syst.*, vol. 13, no. 1, pp. 38–68, 1995.

[23] M. Yakici, M. Baillie, I. Ruthven, and F. Crestani, “Modelling epistemic uncertainty in IR evaluation,” in *SIGIR ’07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2007, pp. 769–770.

Dialogue - Driven Information Retrieval

Adindla Suma
School of Computer Science and Electronic Engineering
University of Essex
Wivenhoe Park,Colchester,C043SQ,UK
sadind@essex.ac.uk

Abstract

Web search engines are built for helping web users to locate information quickly and efficiently. However, despite the fact that search engines provide considerable assistance in locating information, one of the main difficulties remains the ambiguity and uncertainty involved in matching information needs against documents which might satisfy those needs. A possible solution is the application of natural language dialogue systems, an area that is becoming increasingly prominent in the field of natural language processing. Dialogue Systems aim at conversing with a human using a logical and articulate structure. Dialogue systems have been shown to work well over structured knowledge sources. Imposing a dialogue system on an intranet however is a new challenge. Here we are looking at combining a dialogue system with the power of a standard search engine.

Keywords: information retrieval systems, dialogue systems, question answering systems

1. INTRODUCTION

Imagine you could interact with a university intranet search engine as follows:

User: Head of department
System: Which department are you looking for?
User: computer science
System: The head of the computer science department is Dr. Sam Steel. His contact details are as follows ... Do you want any further information?
User: How do I get to his office?
System: The quickest route from the University information point is as follows ...

This type of interaction seems well beyond what is currently possible but the documents stored at the website do contain a lot of implicit structure that could be used to make such a system reality.

One thing we should point out here is that a user is not forced to interact with the dialogue system as there is no need to initiate a separate dialogue to satisfy every user information need. Search engines can perform very well with user queries in most of the cases. Therefore, with any response from the system the user will also see the best matching documents returned by the local search engine. A user is free to ignore all options proposed by the dialogue manager if the top matches contain the desired information.

However, even queries in document collections of limited size often return a large number of documents, many of them not relevant to the query (Kruschwitz, 2005). To stick to the example of the Essex[1] intranet, for instance, the *head of the computer science department* query, the search engine should be able to provide an excellent match to this simple query and retrieve *Sam Steel's* home page as the best match. However, we end up getting several other pages that are not relevant to this query.

One way of imposing a dialogue system on such a document collection is to enforce formatting guidelines on web site developers so that all documents are properly structured, syntactically as

[1] http://www.essex.ac.uk

Parameters	Dialogue systems	QA systems	Intranet domain
Input	natural language questions	natural language questions	short queries(2-3 words) that do not convey much information
Output	very domain specific answers	expected answer type who : person when : date where : location	wide range of queries like room numbers phone numbers course titles
Knowledge base	rely on structured database	free text(search engines) web documents structured databases	unstructured data found on web pages & semi-structured
System-user interaction format	domain specific dialogue manager	either one-shot queries or some element of dialogue	typically one-shot queries

TABLE 1: Related research areas

well as semantically. There are a number of problems with such an approach (Hawking and Zobel, 2007). We do not impose such restrictions and assume that most of the content is unstructured, or partially structured. Our aim is to automatically extract useful domain knowledge from the documents. This knowledege can then be used to guide the dialogue manager. For example, we are interested in extracting subject/object relations to construct domain specific knowledge. This domain knowledge would be quite helpful, to provide a user precise and straight forward results.

The work proposed here represents a PhD project that has just started.

2. MOTIVATION

There has been little progress in the implementation of such dialogue systems in intranet search engines. Our work is based on UKSearch (Kruschwitz et al., 2008), a domain specific dialogue search system developed at the University of Essex. In this, a domain model has been constructed by extracting document markup structure. A user query is submitted to search engine as well as to the domain model. Apart from showing the standard search engine results to the user, the system assists the users by suggesting query modification terms to refine and relax the query. However, the system has no ”understanding” of these modification suggestions.

Our motivation emanated from the user queries submitted to the UKSearch system (Essex University website). Queries have been collected for more than two years to study the user search behaviour. Similar to other studies (Bendersky and Croft, 2009), queries were found to be very short, less than two words on an average. In addition, it was found that queries like room numbers, lab numbers, telephone numbers and course titles etc. were routinely searched for.

3. RELATED WORK

Our research work is closely related to areas like dialogue systems, question answering systems and information extraction etc. Dialogue systems engage in some sort of conversation with a user, to perform some domain specific task. The domain related task could be booking a flight, accessing yellow pages directory data and enquiring about train time tables etc. All these tasks can be categorized as information seeking tasks. The idea here is, users seek some information by explicitly providing some constraints to the system. Table 1 gives a quick overview of the related research areas.

3.1. Information Seeking Dialogue Systems

In recent years, the area of NLP has witnessed a rapid development of dialogue systems from simple conversational agents to sophisticated Multi-modal dialogue systems. Such systems can use various modalities to interact with a user like text,speech,graphics and gestures. We will not look at multi-modal dialogue systems. Early dialogue systems like ELIZA and PARRY are known as conversational agents. The main purpose of ELIZA (Weizenbaum, 1966) was to study the interaction between human and computer, where ELIZA played the role of a psychotherapist. It is a script based program (decomposition and reassembly rules) and reads the input string to identify the necessary keywords. The keyword related transformation rules are applied to the same sentence as a system's response. Although, conversational agents are part of dialogue systems, we are not interested in such type of applications. Instead, we are particularly looking at information seeking aspect in dialogue systems.

Initially, dialogue systems concentrated on travel domain related applications (ATIS) (Hemphill et al., 1990). Later, it moved to other domains like call routing (HMIHY) (Gorin et al., 1997) and intelligent tutoring systems (ITSPOKE) (Litman and Silliman, 2004). Some examples of Multi-modal dialogue systems are: bathroom designer (COMIC) (Catizone et al., 2003), Queens Communicator (O'Neill et al., 2003) and unmanned robot helicopter (WITAS) (Lemon et al., 2001) etc. Unlike these dialogue systems which focus on structured databases, our work concentrates on unstructured data found on the webpages. The significant difference between these dialogue systems and intranets is firstly, we have web type queries. Secondly, queries are more focused in these dialogue systems and finally, we don't have explicit domain specific knowledge.

3.2. Dialogue System Modelling Approaches

Finite grammars are the simplistic approaches to model dialogue systems. Initially, finite grammars were used to model the entire structure of conversation. In finite grammars, states represent system utterance and the transition between the states is based on user's response. The problem with this approach is flexibility and portability issues to other domains (Wilks et al., 2006).

The frame based approaches are an extension of finite grammars. They are also known as slot and filler structures. The slots are filled with in a frame based on user's utterance. The best feature about this approach is multiple numbers of slots can be filled in random order. The database is queried once it has filled all the slots. It is somehow flexible when compared to finite grammars because this approach doesn't enforce any strict ordering on user utterance (De Roeck et al., 1998).

3.3. Question Answering Systems

Question answering (QA) systems automatically extract answers to questions formulated by a user in a natural language. Question answering is not a new research topic and the main goal of these systems is to provide the user short answers, instead of a list of documents. Question answering systems can be broadly classified into two domains:

- Closed domain (limited to particular domain)
- Open domain (can be anything and mostly fact based)

The first question answering systems are Baseball and LUNAR. Baseball answered questions pertaining to baseball games played in the American league over one period (Green et al., 1961). LUNAR focused on questions related to moon rocks and soil collected from the Apollo moon missions (Woods, 1973). These two systems are good examples of closed domain question answering systems. In closed domain question answering systems, the domain specific knowledge is stored in databases. A natural language interface is provided to access that domain information.

In open domain, question answering systems can be divided into factoid and complex QA systems. Most of the open domain question answering systems are factoid based QA systems. Factoid based QA systems deal with facts related to person names, date of births and organizations etc. Factoid question types are classified into: where, when, who etc. On the other hand, for complex and analytical tasks interactive systems like HITIQA (Small et al., 2003) and

FERRET (Hickl et al., 2006) are also developed. Both of these systems are especially meant for answering explanatory questions like: why , how, list , define etc.

QuASM, question answering system depends on html markup structures to extract answers to factoid questions (Pinto et al., 2002). AQUA, (Vargas-Vera et al., 2003) question answering system incorporates various knowledge sources like ontology and WordNet along with a domain specific database. Questions are answered by means of knowledge database. If AQUA fails to answer it uses a search engine to find an answer to a query. We are planning to implement a similar strategy.

Another example of open domain question answering systems is AnswerBus[2]. It considers the user query in any of these languages: English, German, French, Spanish, Italian, and Portuguese and returns results in English. Alta Vista's language translator tool *Babel Fish* is used for translation. Search engines and directories are deployed to find relevant documents to user queries. The major difference with other open domain QA systems is AnswerBus returns a set of possible sentences instead of fixed length answers (Zheng, 2002).

Perhaps the most popular and first web based natural language QA system available online today is START[3]. START (Katz and Lin, 2002) relies on Omnibase, (Katz et al., 2002) a heterogeneous database in which documents are annotated in a natural language to extract triples (subject, relation, and object). In our proposed work, we will also consider these triples to extract domain specific knowledge.

4. RESEARCH QUESTIONS

The questions to be addressed by this proposed research are :

- How can a dialogue system be incorporated into the search process to provide the user with a more natural language interface?
- Can such a dialogue system on an intranet provide the user with more relevant information and offer a better user experience than a standard search engine?

5. PROPOSED RESEARCH

There is a wealth of literature on research in dialogue systems and question answering systems. Dialogue systems do however, typically rely on some sort of structured knowledge whereas question answering systems are in most cases one shot interactions. The proposed dialogue systems for intranet search are different to the related work discussed. The main differences with other areas are :

- Short queries in most cases (often just keywords rather than questions)
- Unstructured data to start with
- Domain-specific dialogue (without having lots of domain knowledge)
- Wide range of possible user queries

The past few years have also seen a rapid explosion of activities in information extraction (Jurafsky and Martin, 2008). Extracting named entities and simple relations has become much more robust and this is another area of research that we will tap into. There are two main parts of proposed research.

- Turn the document collection into a structured knowledge source employing NLP techniques and methods of information extraction to automatically detect named entities like person, names, room numbers etc. As well as facts eg. Predicate - argument structures like:
  is(Sam steel, head of department)

[2] http://www.answerbus.com/
[3] http://start.csail.mit.edu/

- Impose a dialogue system that employs the automatically extracted knowledge and appropriate domain-specific knowledge to assist a user in the navigation as outlined in the motivating example.

Evaluating such a system is particularly difficult as standard measures like precision and recall are not necessarily the best (and certainly not the only) measures to assess a dialogue system. We will perform a range of evaluations, ranging from technical evaluations that investigate the quality of the extracted facts to full user evaluations. Our main methodology to evaluate the proposed dialogue system will be task-based evaluations. Some of the measures that will be used to compare the dialogue system against some baselines (e.g. a standard search engine) include:

- Number of interaction steps required to arrive at an answer (dialogue length)
- Time taken to process a user query
- Precision and recall (success rate of retrieved results)
- User satisfaction

6. ACKNOWLEDGEMENTS

I would like to thank my supervisor Udo Kruschwitz for his help and advice on drafting this paper and three anonymous reviewers for very helpful feedback on an earlier version of this document.

References

Bendersky, M. and Croft, B. (2009), Analysis of long queries in a large scale search log, *in* 'WSCD '09: Proceedings of the 2009 workshop on Web Search Click Data', ACM, New York, NY, USA, pp. 8–14.

Catizone, R., Setzer, A. and Wilks, Y. (2003), 'Multimodal Dialogue Management in the COMIC Project', *Workshop on 'Dialogue Systems: interaction, adaptation and styles of management', European Chapter of the Association for Computational Linguistics(EACL)* .

De Roeck, A., Kruschwitz, U., Neal, P., Scott, P., Steel, S., Turner, R. and Webb, N. (1998), 'The YPA:An Asssistant for Classified Directory Enquiries', *BT Technology Journal* pp. 145–155.

Gorin, A. L., Parker, B. A., Sachs, R. M. and Wilpon, J. G. (1997), 'HOW MAY I HELP YOU', *Speech Communication* **23**, 113–127.

Green, Jr., B. F., Wolf, A. K., Chomsky, C. and Laughery, K. (1961), BASEBALL: AN AUTOMATIC QUESTION-ANSWERER, *in* 'IRE-AIEE-ACM '61 (Western): Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference', pp. 219–224.

Hawking, D. and Zobel, J. (2007), 'Does Topic Metadata Help With Web Search?', *Journal of the American Society for Information Science and Technology* **58**(5), 613–628.

Hemphill, C. T., Godfrey, J. and Doddington, G. R. (1990), The ATIS Spoken Language Systems Pilot Corpus, *in* 'HLT '90:Proceedings of the workshop on Speech and Natural Language', Association for Computational Linguistics, Morristown, NJ, USA, pp. 96–101.

Hickl, A., Wang, P., Lehmann, J. and Harabagiu, S. (2006), FERRET: Interactive Question-Answering For Real-World Environments, *in* 'Proceedings of the COLING/ACL on Interactive presentation sessions', Association for Computational Linguistics, Morristown, NJ, USA, pp. 25–28.

Jurafsky, D. and Martin, J. H. (2008), *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, second edn, Prentice Hall Series.

Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J., Marton, G., McFarland, A. J. and Temelkuran, B. (2002), Omnibase: Uniform Access to Heterogeneous Data for Question Answering, *in* 'Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems NLDB', Springer-Verlag, London, UK, pp. 230–234.

Katz, B. and Lin, J. (2002), Annotating the semantic web using natural language, *in* 'NLPXML '02: Proceedings of the 2nd workshop on NLP and XML', Association for Computational Linguistics, Morristown, NJ, USA, pp. 1–8.

Kruschwitz, U. (2005), *Intelligent Document Retrieval*, Springer, Netherlands.

Kruschwitz, U., Webb, N. and Sutcliffe, R. (2008), 'Handbook of research on web log analysis', *Query Log Analysis for Adaptive Dialogue-Driven Search* pp. 389–416.

Lemon, O., Bracy, A., Gruenstein, A. and Peters, S. (2001), The WITAS Multi-Modal Dialogue System I, *in* 'Proceedings of EuroSpeech 2001', pp. 1559–1562.

Litman, D. J. and Silliman, S. (2004), ITSPOKE: An Intelligent Tutoring Spoken Dialogue System, *in* 'Proceedings of HLT/NAACL', pp. 233–236.

O'Neill, I., Hanna, P., Liu, X. and McTear, M. (2003), The Queen's Communicator: An Object-Oriented Dialogue Manager, *in* 'EUROSPEECH', Geneva,Switzerland, pp. 593–596.

Pinto, D., Branstein, M., Coleman, R., Bruce Croft, W., King, M., Li, W. and Wei, X. (2002), QuASM: A System for Question Answering Using Semi-Structured Data, *in* 'Proceedings of the Joint Conference on Digital Libraries (JCDL) 2002', pp. 46–55.

Small, S., Liu, T., Shimizu, N. and Strzalkowski, T. (2003), HITIQA: An Interactive Question Answering System, *in* 'Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question'.

Vargas-Vera, M., Motta, E. and Domingue, J. (2003), AQUA: An Ontology-Driven Question Answering System, *in* 'Stanford University', AAAI Press, pp. 24–26.

Weizenbaum, J. (1966), 'ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine', *Communications of the Association for Computing Machinery* **9**(1), 36–45.

Wilks, Y., Catizone, R. and Turunen, M. (2006), 'Dialogue Management', *COMPANIONS Consortium: State Of The Art Papers* .

Woods, W. A. (1973), Progress in natural language understanding: an application to lunar geology, *in* 'AFIPS '73: Proceedings of the June 4-8, 1973, national computer conference and exposition', ACM, New York, NY, USA, pp. 441–450.

Zheng, Z. (2002), AnswerBus Question Answering System, *in* 'Proceedings of the second international conference on Human Language Technology Research', Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 399–404.

Lexical Issues of a Syntactic Approach to Interactive Patent Retrieval

Eva D'hondt
Centre for Language and Speech Technology,
Radboud University Nijmegen, The Netherlands
e.dhondt@let.ru.nl

Abstract

Patent retrieval is an information retrieval task that poses very specific characteristics and demands. Especially the need for high recall is very important to patent searchers. In the ongoing research project TM4IP, we aim to improve patent retrieval by developing an open-domain patent retrieval system based on linguistic knowledge. By using Dependency Triplets as index terms our system aims to improve precision and recall compared to keyword-based approaches. One of the cornerstones of a syntactic approach to Information Retrieval is normalisation. This paper describes some of the characteristics of the patent domain that influence lexical normalisation.

Keywords: Patent Retrieval, Natural Language Processing, Dependency Triples, Lexicon

1. INTRODUCTION

Over the last few decades, the increased access to patents and patent-related information has seriously changed the patent world. National patent offices' databases have become available online and the advent of machine translation and OCR technology has enabled patent searchers to widen their search areas and –at least in theory– the effectiveness of their search. *'However, the complexity and concern of getting it right is much higher too ... [among patent searchers, there is] a growing anxiety about missing something.* [1]' The economic repercussions may be vast: for example, a missed patent in a prior art search can lead to an infringement suit, which can cost millions of dollars. While in the past retrieval of patents and other forms of Intellectual Property was typically researched by the database community, more recently it has attracted the attention of the IR community. Patent retrieval has been the topic of workshops in [SIGIR 2000; ACL 2003; NTCIR3 2002; NTCIR4 2004] and this year the CLEF-IP 2009 track is directed exclusively at prior art search.

Patent retrieval is substantially different from ad hoc retrieval, because of the special characteristics of both the documents and the queries (as well as the goals of the searchers) that are involved in the search.

Depending on the specific purpose for undertaking the search, the end point of patent searches starting from the same query may be quite different. Prior art search, invalidity search, infringement watch or state-of-the-art search all have different goals and information needs and the relevance of the information in the set of found documents will be evaluated accordingly. This shows the need for a patent retrieval system to have a well-developed, interactive search component which can be modified to suit the goal and taste of the searcher.

When composing their search queries, patent searchers are very much concerned with the fact that they cannot afford to miss an important patent. I have dubbed this the 'total recall problem': The patent searcher is willing to lose precision in an effort to reach maximum recall. Expert users create long Boolean queries (comprising 5 to 30 terms) where each concept searched for is expressed by AND's and OR's of possible synonyms [3]. The resulting set of documents is then analysed document per document to judge their relevance. The exposure to alternative descriptions of some concept can lead to a careful rephrasing of the original query and so the process is repeated until either the relevant information is found or the (known) list of variations is exhausted. This way of searching is very labour-intensive: some search sessions can take up to two months to locate 200 relevant patents [3]. There have been some attempts to facilitate this search process by automatically extending the search queries with synonyms or ontologically related terms, but such systems have not met great enthusiasm from the patent searchers community. As [5] point out:

> ’The professional searcher does not like the black box effect of intelligent engines, for example the automatic query expansion with synonyms, stemming, default use of the OR Boolean operator, etc. Whether a box is really black or just looks black because of absence of illumination (in this case knowledge of the linguistic algorithms used) can become the subject of philosophical speculations.’

While patent searchers realize that they need help to achieve the highest possible recall, they also demand total control and constant insight in precision and recall. Any retrieval system aimed at this specific task and public should be designed with these two concepts in mind: ’total recall’ and ’total transparency towards the user’.

If we turn to the patent documents another set of problems appears:

Patent texts are not homogeneous, but consist of different sections[1] which use different styles. However, depending on the search goal the data relevant to (part of) the information need may be in any of these sections; so all of them must be taken into account. The claims section, for example, is legally bound to be one massive sentence. This is where the legal protection of the invention is determined, which results in even more legalese than usual.

Even in the more descriptive sections, the language of patent documents is still notoriously difficult to read. This is partly because of their grammatical structure: Patent texts often contain long, complex sentences with a lot of enumerations and ellipses. However, as argued by [8], in most patent texts, the grammar is correct and only a subset of the grammatical constructions in a language are used. This allows for an efficient analysis by a (specially developed) parser. Yet a greater challenge of processing patent texts, more specifically from an IR point of view, is the legal style in which patents are written. ’Patentese’ combines the specific vocabulary of the domain to which the patent applies with very generic terms and expressions. (The latter is a tactic that is commonly used by patent lawyers who aim to obfuscate the method and specifics of the invention.) This huge variation in expressing concepts makes it very difficult to find all relevant documents pertaining to one’s information need. As [1] puts it: *’In the patent world, words are shapeshifters, yet words are the brick and mortar of modern information retrieval.’*

The keyword-based systems that make up the majority of todays patent retrieval systems cannot look past this variation. A retrieval method based solely on surface forms encounters all sorts of problems: The difference between the noun and verb ’means’ is undetectable for a pure keyword-based approach. A second problem encountered by these systems is the fact that they cannot deal with the morphological variation of word forms. Even in a keyword-based search which incorporates stemming, the fact that ’index’ and ’indices’ are basically the same word will not be noticed. Nor can keyword-based retrieval deal with the similarity between ’adhere’ and ’adhesion’ (nominalization). Because of the fundamental limitations of keyword search we have to go beyond the surface form. Therefore, we propose a syntactic approach to patent retrieval in which the syntactic (and correlated semantic) relations [10] between two words are the index terms. We use Dependency Triplets like `[N:candle, SUBJ, V:burn]` to represent different but equivalent expressions like ’the burning candle’, ’the candle(s) burned’, (he was) burned by a candle’, etc . By doing so we can abstract away to a level beyond surface realization, closer to concepts thus resolving problems of ambiguity due to morphological and syntactic variation.

[1] title, abstract, description and claims section

My PhD research project is part of the TM4IP[2] project [4] which aims to develop an open-domain patent retrieval system based on linguistic knowledge. My research focuses on the linguistic concerns that arise while designing and implementing the parser and search module. Until now, the syntax and semantics of patent texts have received little attention from the linguistic point of view. While there are a multitude of problems and considerations that are relevant for linguistically informed patent search, the focus of this paper is more specifically on the lexical characteristics of patent texts which influence the process of transforming this type of text (genre) into usable/optimal index terms.

In section 2, I will briefly describe the state of the art in patent retrieval systems and give a brief overview of the system used in the TM4IP project. In section 3, I will zoom in on an experiment that I performed and discuss plans for future research.

2. BACKGROUND

2.1. Patent retrieval systems

The majority of the search engines used by the patent search community today are keyword-based, using a general-purpose text search engine. Some of them incorporate a query preprocessing module and allow for the use of wild cards, boolean compositionality and term weighting [12], query phrases, query expansion by using thesaurus [12], proximity searchc̃itemicropat, etc. For an overview of the three biggest commercial systems, see [11]. The vector space model usually lies at the heart of these generic keyword-based approaches. Academic research has mainly focussed on the relative weighing of these terms [6] and on exploiting the patent document structure to boost retrieval efficiency [6]. Over the last few years the first semantically based patent retrieval systems have been introduced, in particular the Patent-café search engine and IPCenturys DECOPA search engine.

The only method that comes close to our syntactic approach is [9] who uses deep linguistic analysis in the form of predicate-argument analysis (implying semantic role labelling) to improve readability. Her system is the first step in a suggested patent summarization method. A closely related system is the PATexpert system [2], a content-oriented system that aims to look past the surface forms and uses –amongst other technology like image retrieval, etc– semantic web technology to give direct access to the content of the patent. Due to the highly specialised ontologies used, the PATexpert system is currently (2008) focused on two domains: optical recording devices and machine tools. In TM4IP we want to avoid this dependence on elaborate ontologies and have opted for an open-domain system.

2.2. Introducing the PHASAR system in the TM4IP project

The basic unit in our system is the Dependency Triplet (DT). A DT is similar to a syntactic phrase in that it is a grammatical part of the sentence and –at least in part– identified according to linguistic criteria. The use of syntactic phrases in Information Retrieval is based on the assumption that words in a text that have a syntactic relationship often have a related semantic relationship [10]. A DT is a pair of (lemmatized) words together with their syntactic relation, e.g. `[N:current, ATTR, A:electrical]`. PHASAR's DT framework is based on the principle of aboutness.

Our system is made up of two main components:
a) AEGIR (Accurate English Grammar for Information Retrieval), a hybrid dependency parser, which aims to accurately parse complicated technical English texts for IR purposes. AEGIR combines a broad-coverage, handcrafted rule-based grammar with a transduction process to (a) find the best parse and (b) transduce this parse to DTs while processing large raw corpora in the patent domain. Then, the information about the frequency of DTs and lexical items is incorporated into the parser, thereby guiding the parsing process as it analyses new text. For example, the correct parse of the famous example John hit a man with a telescope could easily

[2] *'Text Mining for Intellectual Property.'* For more information on the project, visit http://www.phasar.cs.ru.nl/TM4IP.html

be decided given the information that `[V:hit, PREPwith, N:telescope]` occurs 2 times in the entire (training) corpus, while `[N:man, PREPwith, N:telescope]` occurs 220 times.

b) PHASAR (Phrase-based Accurate Search And Retrieval) is a search engine which uses DTs as index terms. In the search module, the searcher can phrase queries in a semi-natural way to fit the index terms as closely as possible. One has complete control over the search and search result and can interactively generalize the query or make it more specific. Query generalization can be achieved by either joining multiple terms using the OR operator, or by using one of the built-in thesauri for selecting a semantic term type. A query can be made more specific, by adding more terms in the query slots or by setting a context from which the results have to be retrieved.

In our system, the normalisation needed to create effective index terms is taken care of by different parts of the system.

- **Grammatical normalisation** is built into the parser, which has a second transduction step that translates a dependency graph into a set of DTs. Special care has been taken to reach the highest grade of normalisation possible, for example by splitting up hyphenated forms such as 'man-made lake' into `[N:man, SUBJ, V:make]` `[V:make, OBJ, N:lake]`, to map nominalisations onto the relevant verbs, to map equivalent grammatical structures onto each other, e.g. 'thumb movement' onto 'movement of thumb'.
- **Morphological normalisation** is realised partly through the lexicon and partly through the parser. Spelling variation and morphological variations such as singular-plural, tense or mood are handled by abstracting to a lemma (present in the lexicon). For those forms that have no entry in the lexicon, robust recognition rules are incorporated in the system.
- **Semantic normalisation** is (partly) realised by using the DTs themselves, which provide a disambiguating context (achieving higher precision). E.g. tree bark versus a dog that barks.
- **Lexical normalisation** is (partly) realised by the use of (several) thesauri which can be accessed by the user during a search and analysis phase.

3. CONSIDERATIONS AND POSSIBLE RESEARCH ISSUES

As explained in the previous section, problems of syntactic, morphologic and -to some extent– semantic variation are dealt with by the parser. But one problem still stands in the way of maximum recall: lexical variation. While being a big problem in any IR task, there are some characteristics of patent texts that make lexical variation even more difficult to deal with in the patent domain.

3.1. Diversity in the patent domain

The notion of the 'patent domain' is a deceptive one. The so-called patent domain actually consists of hundreds of highly technical and specialised subdomains, each with its own very specific terminology and ontologies. A patent document may concern anything from a gene extraction method, a business method or chemical compound to a particular design of a doll house.

With such a wide variety of specific subdomains, finding and extending reliable lexical resources is very difficult. Yet, ontologies and lexica are very important for all parts of our system. They are not only used for query extension or classification tasks, but the lexicon is the very basis of our parser system. When the parser encounters a word in the text that it cannot find in its lexicon, it can either ignore it, thus -unlike the bag-of-words approach– creating 'gaps' in the information processing of the text, or use a series of robust recognition rules to make an informed guess as to the relation of this lexical item compared to the other item in the dependency triplet. The latter approach is more error prone and often does not provide the correct POS information or lemma, thus inserting noise into the set of index triples.

3.2. Lexical and semantic variation

Even if we could have access to perfectly-constructed and complete ontologies, a second defining characteristic of the patent domain would still pose a serious problem: semantic variation or the possibility of one expression to denote several concepts. The patent domain exist through the collaboration of (tens of) thousand of writers and as there is no-agreed upon vocabulary (like for example in the UMLS), every inventor has the right to use a word as he or she sees fit. For example, a multiword term such as 'sound emitting device' can denote everything from a car horn or headphones to the so-called Mosquito Device, a device which is designed to drive young troublemakers away from a problem area. This makes it very difficult to perform an effective keyword search/direct term-matching To deal with this problem we need to disambiguate the meaning of the words that are encountered in the patent text.

Closely related to semantic variation is the problem of lexical variation: the same concept can be expressed in different patent texts with different lexical items, e.g. 'spring', 'wire of coiled steel', 'means compressible along an axis', etc . All these synonymous expressions should be identified as such and linked to each other by means of a thesaurus or ontology.

3.3. Vague terms

A third characteristic of patent texts is the use of vague terms and expressions. Terms such as 'sound emitting device', 'means compressible along an axis' or 'apparatus for preparing and dispensing whipped beverages' are used instead of 'loudspeaker', 'spring' or 'drink dispenser'. As mentioned above, this kind of lexical variation complicates direct term-matching. These terms are invented and defined ad hoc by the patent writers and will generally not end up in any dictionary or lexicon. Exceptions are terms whose acronyms have become common words, like LED (light emitting device) or LCD (liquid crystal display). Our system -at present– cannot catch these terms and will normalize them into separate DTs, essentially treating the information as if it would occur in a regular sentence instead of appearing as a term.[3]

While this syntactic normalisation usually improves recall [7], treating the information in sentences and inside terms in the same way has disadvantages for lexical normalisation. If a patent searcher is looking for all patents concerning headphones, he will also be interested in an obscure patent concerning a 'sound emitting device'. Instead of only splitting the information up in separate DTs, the system should be able to use these terms for the lexical normalisation process as well by linking them to known terms. This is equivalent to expanding the systems ontology with these ad-hoc synonyms. This requires a two-step approach: correctly identifying and extracting vague terms and linking them to an existing ontology.

Preliminary analysis of 16,000 patents in the engineering domain revealed that vague terms like the examples above are headed by so-called general-purpose words[4]. These words are semantically light: they occur quite often in the corpus (see appendix 2 for the occurrence of general-purpose terms in the 20 most frequent terms) but are, by themselves, not very informative. Their meaning is rather abstract and they are always accompanied by a verb or an adjective derived from a verb that specifies their function. I created a list of general-purpose words by looking at the 200 most frequent words and selecting the words which fit the criterion of 'whilst being an instrument or method the word does not contain an intrinsic expression of its function'. This was done by 2 persons with an inter-annotator agreement of 67%.[5] The resulting list can be found in appendix 1. I then looked at the noun phrases which where headed by these words.

[3] The three examples given above would respectively become `[N:device, SUBJ, V:emit], [V:emit, OBJ, N: sound]; [V: compress, OBJ, N:means] , [V:compress, PREPalong, N:axis]` and `[N:apparatus, SUBJ, V:prepare], [N:apparatus, SUBJ V:dispense], [V:prepare, OBJ, N:beverage], [V:dispense, OBJ, N:beverage], [V:whip, OBJ N:beverage]`. In this paper I do not discuss the more typical compound terms (a noun phrase in which the adjective or noun is attributively connected to the head noun), e.g. thermotherapeutic apparatus or liquid crystal display device. While these compounds are equally important in the search for lexical normalisation and will receive a great deal of attention in my research project, they are less specific for the particular difficulties of the patent domain than the 'vague terms'.

[4] A general-purpose word can be defined as noun that describes an article or instrument whose function is not expressed by the word itself but through the context words. For example, 'apparatus' in the term 'occupant sensing apparatus' does not express its function, while the word 'sensor' does.

[5] Inter-annotator agreement was calculated by counting each annotation of person 2 as a match or nonmatch value to the annotations of person 1 and then calculating the ratio of matches to the total number of annotations.

Manual analysis of the noun phrases showed that these vague terms usually follow one of these syntactic formats:

- Noun Phrase – Present Participle – General-Purpose Noun, e.g. 'cover folding device' or 'digital video/audio recording and reproducing apparatus'
- General-Purpose Noun – Adjective or Adjectival Phrase, e.g. 'means compressible along an axis',
- General-Purpose Noun – for/of – Present Participle Noun Phrase, e.g. 'device for making sandwiches', 'apparatus for dispensing medicine', 'a means and method for implanting bioprosthetic material'.

In a few patents small variations occurred within the vague terms. For example, a patent describing a 'cover folding device' would sometimes use the term 'top cover folding device' instead. These variations are a sign of the ad-hoc status of these terms.

Analysis of their distribution in the patent documents shows that these vague terms appear frequently in specific places such as the title, the abstract and the claims section. The description section of the document deals with the parts and components of the system, so naturally such general terms do not occur very frequently in this section. What is interesting, however, is the low frequency of these terms in the prior art descriptions. While one would expect relatively high frequency as they are describing similar concepts patent writers tend to opt for more concrete terms to denote the same concept. For example, the patent discussing an 'occupant sensing apparatus' would use the term 'sensor' in the prior art description, but not in the rest of the document.

The next step of my project will be to design a system that can a) automatically extract vague terms and b) link them to existing ontologies. One of the challenges of automatic term extraction will be to decide which words belong to the n-gram. As we are looking for relatively big terms (more than 3 words) the complexity of the task increases enormously. Next to our knowledge of the general-purpose terms list and the syntactic templates, one characteristic of patent documents that may help solve this problem is the patent writer's need to avoid ambiguity. As mentioned above there is a great deal of intertextual variation in the expression of concepts even within one domain. Yet here is almost no lexical variation in the patent texts themselves. Since failure to describe an invention or claim clearly and unambiguously can result in rejection during prosecution [1], the patent lawyer will not risk any misunderstanding and thus will always refer to a concept that has already been introduced in the text by means of clear anaphoric elements or will just repeat the term in full. In due time we will develop an anaphora resolution module, but for now the system can only look at those instances where the whole vague term was repeated. Using the linguistic knowledge of the parser to detect the three syntactic templates described above we can detect the potential vague terms. If such a term appears in the title, abstract and/or frequently[6] in the claims section of the patent document, the system would classify it as a 'vague term'.

In subsequent work, the automatically extracted vague terms should be used to expand existing ontologies. We would like to know if a vague term could be a synonym for a more concrete term that is part of the ontology. As mentioned above, in the prior art description the patent writer often uses more concrete terms than in the rest of the patent to describe the same or a very similar concept or invention[7]. My hypothesis is that the verb in the vague term is a very good indicator of the desired (more concrete) term. It describes the function of the vague term and often occurs (almost) literally in the concrete term. For example, an 'occupant *sensing* apparatus' could easily be linked to 'occupant *sensor*' (using the nominalisation database NOMLEX). I will investigate if focusing on the term (and its synonyms in WordNet) can help identify less obvious concrete terms in the prior art section.

[6] That is, appear more often than a not yet specified cut-off ratio, normalised to the length of the claims section and abstract.

[7] For example, a patent concerning a 'device for cooling an infant's brain' in which this term was used very consistently used 'cooling cap' in the prior art section.

4. CONCLUSION AND DISCUSSION

In this paper I presented an overview of the characteristics of patent retrieval and discussed some of the lexical issues that are particular to the patent domain. These issues where described in the context of the TM4IP project, in which we develop a patent retrieval system based on a syntactic approach. The design behind the PHASAR search engine aims to get very high recall by achieving syntactic and morphological normalisation. We are currently looking for ways to extend our approach by adding extra lexical normalisation.

As was shown in section 3 lexical variation is an important problem in patent retrieval. The specific requirements of the patent searchers (i.e. 'total recall' and 'total transparency') demand a novel approach that combines both an automatic extraction of potential new terms and an interactive component allowing the patent searcher to keep control over the search terms in the query. In section 3.3, I sketched an initial implementation but this is by no means complete. Hence, I am open to any advice or comments on the current approach, or on alternatives and extensions of the approach. Specifically, there are a number of open issues that I would like to get feedback on:

- If vague terms can be identified, there still is the challenge of attaching them at the correct positions in the ontologies. What would be a good way to automatically extend existing ontologies?
- Determining which is the correct synonym will prove to be a major challenge. What do you imagine would be the major pitfalls?
- For those who have a background in patent retrieval: there is a common feeling that different sections of patent texts contain very different information, yet precious little has been published on this subject. Could anyone give me some pointers?

REFERENCES

[1] Atkinson K. (2008) Toward a more rational patent search paradigm. In *Proceedings of the 1st ACM Workshop on Patent information Retrieval*, (Napa Valley, California, USA, October 30 - 30, 2008). PaIR '08. ACM, New York, NY, 37-40.

[2] Escorsa E, Giereth M, Kompatsiaris Y, Papadopoulos S, Pianta E, Piella G, Puhlmann I, Rao G, Rotard M, Schoester P, Serafini L and Zervaki V. (2008) Towards content-oriented patent document processing. In *World Patent Information* 30(1):21:33.

[3] Homan H. (2004) Making the Case for Patent Searchers. In *Searcher*, 12(3).

[4] Koster C, Oostdijk N, Verberne S and D'hondt E. (2009) Challenges in Professional Search with PHASAR. In *Proceedings of DIR 2009*, pp. 101-102

[5] Krier M, Zacca F. (2002) Automatic categorisation applications at the European patent office. In *World Patent Information* 24(3):187 96.

[6] Mase H, Matsubayashi T, Ogawa Y, Iwayama M, and Oshio T. (2005) Proposal of two-stage patent retrieval method considering the claim structure. In *ACM Transactions on Asian Language Information Processing* (TALIP) 4, 2, 190-206.

[7] Moens M-F. (2005) *Automatic Indexing and Abstracting of Document Texts.* The Kluwer International Series on Information Retrieval Vol. 6. 343-347.

[8] Sarasua L, Corremans G. (2000) Cross lingual issues in patent retrieval. In *Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval.* ACM, New York.

[9] Sheremetyeva S. (2003) Towards Designing Natural Language Interfaces. In *Proceedings of the 4th International Conference "Computational Linguistics and Intelligent Text Processing"* Mexico City, Mexico

[10] Smeaton A, Sheridan P. (1991) Using morpho-syntactic language analysis in phrase matching. In *Proceedings of RIAO'91.* Barcelona, Spain, pp. 414-430.

[11] http://www.infonortics.com/chemical/ch04/slides/lambert-new.pdf

[12] http://www.delphion.com

[13] http://www.micropat.com

APPENDICES

A. LIST OF GENERAL-PURPOSE TERMS IN THE ENGINEERING DOMAIN

N:system
N:apparatus
N:method
N:control
N:appliance
N:device
N:holder
N:tool
N:implement
N:member
N:body
N:means
N:assembly
N:frame

B. 20 MOST FREQUENT NOUNS IN THE CORPUS (GENERAL-PURPOSE TERMS ARE UNDERLINED)

4759	N:end
4089	N:device
3995	N:portion
3943	N:surface
3918	N:control
3448	N:position
3362	N:output
3160	N:system
3082	N:member
3036	N:apparatus
2996	N:current
2978	N:means
2919	N:frame
2737	N:material
2654	N:assembly
2650	N:air
2634	N:circuit
2566	N:support
2483	N:tool
2393	N:cylinder

Multimodal Information Spaces for Content-based Image Retrieval

Juan C. Caicedo
National University of Colombia
http://www.informed.unal.edu.co
jccaicedoru@unal.edu.co

Abstract

Currently, image retrieval by content is a research problem of great interest in academia and the industry, due to the large collections of images available in different contexts. One of the main challenges to develop effective image retrieval systems is the automatic identification of semantic image contents. This research proposal aims to design a model for image retrieval able to take advantage of different data sources, i.e. using multimodal information, to improve the response of an image retrieval system. In particular two data modalities associated to contents and context of images are considered in this proposal: visual features and unstructured text annotations. The proposed framework is based on kernel methods that provide two main important advantages over the traditional multimodal approaches: first, the structure of each modality is preserved in a high dimensional feature space, and second, they provide natural ways to fuse feature spaces in a unique information space. This document presents the research agenda to build a Multimodal Information Space for searching images by content.

1. INTRODUCTION

Content-Based Image Retrieval (CBIR) is an active research discipline focused on computational strategies to search for relevant images based on visual content analysis. In this proposal, multimodal analysis is considered to develop CBIR systems, specially for image collections in which there is some text associated to images. Multimodality in Information Retrieval is sometimes referred to the interaction mechanisms and devices used to query the system. However, since the Multimedia Information Retrieval perspective, multimodality is referred to those methods that take advantage of different data modalities to provide access to a digital library or a multimedia collection [11, 5]. Different data modalities in multimedia are used to better understand document contents, including textual annotations, audio, images and video. In this proposal, *multimodal* will refer to the ability to represent, process and analyze two data modalities simultaneously: unstructured texts and images.

In our daily life many multimodal collections composed of unstructured text with associated images may be found, for instance, the web, which is largely composed of pages with text paragraphs and several images. Other examples of document collections with this structure are scholarly articles, book collections, news archives and medical records. The majority of these document collections are accessed nowadays using information retrieval systems devised to index text contents, so that users can search expressing their information need using textual keywords. However, the vast amount of non-text data available in many document collections may lead to the design of other effective ways for accessing and finding information.

Let's take a look to some examples of the previously mentioned multimodal document collections. Imagine you are traveling for the first time to some place and you find an interesting building that caught your attention, but you do not have any information about it. Then, you capture a photograph using your camera phone and send it to a web search service that returns a list of related web pages with historical information, technical data and tour guides [22]. This service would be useful not only for touristic places but also for products, devices and movie posters among others. Let's move to a clinical environment in which you are a physician evaluating a patient with a medical image and, according to your experience, this image has a non-usual appearance. Then, you decide to query the medical information system in order to get a set of similar images evaluated by other physicians [13]. In addition to the similar cases, you obtain as result some recent medical papers also related to the image contents.

Both examples present situations in which the user is not able to express an accurate query using keywords. Instead, the use of the image at hand may prevent a trial-and-error loop using different keyword combinations and may offer a more precise way to access the right information. Moreover, a multimodal document collection may also be accessed using a multimodal query, i.e. a query composed of images and text descriptions. In this case, the query may lead to find highly relevant documents, since text descriptions give semantic and contextual hints about image contents and the image may help to disambiguate the right meaning of the text description [6].

The study of multimodal information retrieval systems is proposed in this document. In particular, the design of computational strategies to take advantage of multimodal interactions between image contents and unstructured text descriptions is proposed to improve the response of an image retrieval system. In addition, the evaluation of different query paradigms is proposed, including query by example, a keyword based and multimodal queries to search for images. A unified framework is proposed in this document to manage data representation, search algorithms and query resolution. The study and evaluation of kernel methods to generate Multimodal Information Spaces is proposed.

This proposal aims to approach practical and theoretical aspects of a multimodal information representation for image retrieval systems. The proposal is based on kernel methods, which provide foundations to include structure in data representation and also to combine different heterogeneous data sources. Kernel methods for pattern analysis have been studied to design machine learning algorithms, and have been widely used for non-vectorial data, such as strings, trees and graphs among others [17]. Adapting such a framework for information retrieval, and specially for multimodal information retrieval may lead to more effective systems, and also may contribute to the understanding of the relationships between information retrieval and machine learning.

2. PREVIOUS WORKS

The research field on content-based multimedia retrieval has largely grown in the past few years. Datta et al. [5] performed a simple exercise to validate this hypothesis finding a roughly exponential growth in interest in image retrieval during the last 10 years. Video and audio retrieval have attracted great interest too, leading to a more general case of information access into multimedia collections [11]. A pool of events in multimedia information retrieval have been organized to establish common test collections, evaluation protocols and baselines in a competitive environment to academically share research experiences. The following are the set of more prominent events: TRECVid [18] for content-based video retrieval, INEX [8] for structured multimedia collections, ACM-MM GrandChallenge [16] for large multimedia collections and CLEF [3] for image, video, audio and cross language collections. Importantly, most of the defined collections in these events are composed of multimodal data, and recent studies suggest the potential advantage of using multimodal synergies in image databases [5, 11].

The core strategy in Multimodal Information Retrieval is the combination or fusion of different data modalities to expand and complement information. Previously, it is important to process each independent modality. In the case of text documents, well known strategies such as the Vector Space Model are very effective and have been largely extended to solve different problems [12]. The CBIR community does not have a general agreement in the kind of image representation or retrieval model that may be applied. However, recent experimental evaluations are giving a more clear panorama of the representation problem [2], suggesting some promising directions. In fact, some of them are becoming popular in the current research such as the bag of features and statistical signatures.

Once each modality is processed to extract the most informative data, the combination procedure is applied. Two ways to combine multimodal information can be identified: late fusion and early fusion. Late fusion refers to those methods that preserve each data modality separately and, when a user request is received, two search algorithms are executed, one on each modality, to integrate the results just before deliver them to the user. On the other hand, early fusion is referred to those methods that integrate both data modalities before a user request is received, i.e. data have been previously fused and the search algorithm runs on the new fused representation. Both fusion strategies have been subject of recent research and they have provided important insights for the operation of Multimodal Information Retrieval systems.

Late fusion, i.e. combining different rankings, is also referred to as rank aggregation or data fusion [14]. In information retrieval, data fusion merges the retrieval results of multiple systems and aims at achieving a performance better than all systems involved in the process. There are several algorithms to combine rankings that are well known in the information retrieval community, such as linear combination of rankings, summing all similarity scores for each document and voting algorithms inspired in social sciences among others. These algorithms have been evaluated in text information retrieval showing an improved operation [21]. Other simple algorithms based on sets operations to merge ranking lists have been evaluated for image retrieval, using a text search engine and a content-based image retrieval system [19]. In addition, Lau et al. [9] showed that linear combinations of text and visual rankings may lead to better results than each individual system.

Early fusion aims to build an integrated representation of multimodal data to take advantage of implicit relationships. The most simple approach is to normalize and concatenate feature vector representations of each modality. Ayache et al. [1] evaluated this approach to index a video collection with multimodal information. For image retrieval this approach has also been evaluated and extended using Latent Semantic Indexing [15]. An image is considered a document with text data in a vector space model and and visual patterns represented by a bag of features. Both representations are projected together to a latent space in which the search for similar images is performed. Canonical Correlation Analysis (CCA) has also been proposed to find relationships between visual patterns and text descriptions. For instance, Vinokourov et al. [20] applied Kernel CCA to a web image collection to identify links between visual and text representations in order to solve cross modal queries. More recently, the problem of early fusion has been reformulated as a subspace learning problem that offers both dimensionality reduction and feature fusion [7]. The general problem of feature fusion is of great interest in multimedia processing for applications in classification and retrieval tasks.

3. RESEARCH PROBLEM

The main three strategies for Multimodal Information Retrieval are: semantic image retrieval, late multimodal data fusion and early multimodal data fusion. We argue that translation and auto-annotation models for semantic image retrieval lose information of image structure summarizing it into keywords. Then, visual appearance, scene composition and other visual hints are simply discarded. On the other hand, the late fusion approach uses simple strategies to combine the results such as linear combinations or voting algorithms, and the interactions between texts and images may be more complex than that.

The proposed research focuses on strategies for early multimodal data fusion to model interactions between different data modalities. A Multimodal IR system under that approach has three main associated issues as follows:

1. *Content representation of each modality*. The content representation involves the analysis and extraction of information from each modality separately. The processing of image contents and text documents is the main task in this step. It allows the filtering of non-useful data and capture the most discriminative content as is usually done in information retrieval systems.
2. *Information fusion.* The information fusion step, a particular aspect of Multimodal Information Retrieval systems, leads to the design of methods to find and represent the relationships

between both modalities. How to discover the most meaningful associations between images and text and how to complete missing data or non-clear relationships, are the main problems in this step. In this research, the design of early fusion methods is proposed, so at the end of this step, a new document representation is obtained containing both, visual and textual information.

3. *Multimodal retrieval algorithms*. Multimodal retrieval algorithms on the fused representation are designed to identify the most relevant results for the user. The main research questions in this step are related to the query representation and how to solve unimodal and multimodal queries.

4. PROPOSED RESEARCH

The main goal of the proposed research is to design and evaluate a Multimodal Information Space for content-based image retrieval. The construction of such a space is based on kernel methods, that provide a strong theoretical frame to work with different complex and structured data representations. Kernel methods have had a great impact in machine learning and pattern recognition, since they provide effective algorithms and strong theoretical properties. Shawe-Taylor & Cristianini [17] present four principles of a kernel method solution that will be followed in this proposal to approach the problem of Multimodal Information Retrieval systems:

1. *Data items are embedded into a vector space called the feature space.*
2. *Linear relations are sought among the images of the data items in the feature space.*
3. *The algorithms are implemented in such a way that the coordinates of the embedded points are not needed, only their pairwise inner products.*
4. *The pairwise inner products can be computed efficiently directly from the original data items using a kernel function.*

The following subsections present the outline of the main theoretical properties that are considered to tackle the problems of how to represent image and text document contents, how to address the fusion and combination problem using kernels and how to solve queries in a Multimodal Information Space.

4.1. Content Representation

The content representation will follow the first and fourth principles of a kernel method solution. Using the first principle we have a vector space for each data modality in the Multimodal Information Retrieval system. The key point is that the vector space is implicitly defined by the kernel function for the data that is being analyzed. A kernel function gives a similarity notion between the input data. So that, following the fourth principle, we can devise efficient methods to embed the input data into that vector space without explicitly define it. Kernel functions have attracted a lot of attention in different pattern recognition tasks, such as structure prediction in bioinformatics, text categorization and image classification. And since a kernel function provides a mechanism to introduce a similarity measure of two objects in the learning system, many of the proposed algorithms to calculate the kernel value take into account the object structure. In that way, the feature space in which the data is actually represented, is usually a high dimensional vector space that contains information about the structure and the content of the original object.

In the particular case of image and text processing, different kernel functions have been proposed. For instance, the computer vision community has developed some kernel functions to represent images using local appearance and global structure, e.g. the Spatial Pyramid Match Kernel [10]. Other kernel functions for images include histogram kernels, segmentation graph kernels and spectral based kernels among others. On the other hand, for textual documents and strings some kernel functions have also been defined to capture syntactical structure and semantic relationships [17]. Those kernel functions on both, visual and textual data, have shown effectiveness and robustness in challenging classification tasks, obtaining state-of-the-art performance. This suggest that, those kernel functions may have also a good performance in an information retrieval task. Importantly, those functions generate a vector space to represent data of each modality.

4.2. Information Fusion

Information fusion will follow the first and second principles of a kernel method solution. According to the previous subsection, using a kernel function for text and another one for images, we have two vector spaces in which each modality is represented independently. Under a kernel framework, there are different strategies and algorithms to operate in two vector spaces simultaneously depending on the pattern of interest. So for example, if we would like to combine two vector spaces, we can generate a new vector space containing the information and structure of both modalities just by adding the corresponding kernel functions. Imagine a new vector space with structural information of images and semantic meanings of texts. This space may have million of dimensions to represent such contents, the good news is that we do not need to explicitly calculate a matching function between those vectors, instead, operate with kernel functions.

A set of operations with two kernels have been identified such that the resulting function is a valid kernel too [17]. Some of those operations include, addition, multiplication and composition among others. Depending on the operation used to combine two kernel functions, the vector space associated to the new kernel may be of a higher dimensionality. In addition, each dimension in this new vector space may be weighted in different ways. According to the fourth principle of a kernel method solution, the feature space is devised to provide linear relations among data items in the feature space. This also suggest the use of other pattern analysis algorithms to identify more meaningful relationships between images and texts in the feature space.

4.3. Information Retrieval

The Multimodal Information Retrieval algorithms will follow the third principle of a kernel method solution: algorithms only need the inner product information between vectors in the feature space, instead of the explicit vectors. In that way, we can imagine again a vector space with image and text information. If we want to calculate the matching or the similarity between a stored document and a query in the feature space, we can use the kernel functions to do so. Suppose the desired measure to be applied is cosine similarity in the Multimodal Information Space. Since a kernel function is the dot product of two elements in the feature space, we can easily calculate the cosine similarity using the following expression: $cos(x, y) = k(x, y)/\sqrt{k(x, x)k(y, y)}$. It would directly simulate the ranking obtained in a traditional information system if our feature space is the vector space associated with term frequencies and the kernel function is the identity.

Other more sophisticated algorithms may be applied in the Multimodal Information Space, since the mathematical framework of kernel methods has found direct relationships between the Vector Space Model and a feature space. For instance, Latent Semantic Analysis may be applied in the Multimodal Information Space using Latent Semantic Kernels [4], then, we can select a subspace to project the multimodal data in which each dimension stands for a latent semantic topic in the collection. The availability of these tools show that many of the operations that are currently applicable to a Vector Space Model in Information Retrieval can be extended to a feature space, that has been called the Multimodal Information Space through this proposal.

4.4. Evaluation

There are a lot of document collections that include both, images and texts, in which users require to find information either illustrated in images or described in texts. This project has as goal to index the information of images and texts simultaneously to find relevant information independently of its original format. Although the kind of collections on which such a system may be applied is very diverse, this project aims to evaluate the proposed system in a collection of medical information, including images, medical records and scholarly papers. In particular, the collections provided in the ImageCLEFmed competition are planned to be used as well as the datasets collected in the Bioingenium Research Group, product of its operation.

A prototype system will be implemented to operate with the proposed methods, particularly to search for relevant documents given multimodal or unimodal queries. The response of the system will be assessed using standard IR measures to compare results with reported baselines and state-of-the-art methods. The response of the proposed Multimodal Information Retrieval will be also compared with the response of a standard text search engine and a standard image retrieval system to evaluate their relative performance. It is expected that the Multimodal Information Space provide more accurate results.

4.5. Performance Considerations

The proposed research is mainly based on kernel methods that may work on very high dimensional spaces. Kernel based algorithms do not need to operate explicitly in the high dimensional space, and that leads to the implementation of fast similarity measures between structured data. For example, the Pyramid Match Kernel [10], used to approximate the matching between two sets of image features, provides high accuracy and low computational effort compared to the optimal correspondences between the sets' features.

However some learning algorithms need to process a kernel matrix that grows quadratically with the size of the sample. For instance, a Singular Value Decomposition (SVD) of the kernel matrix is useful for doing principal component analysis or latent semantic analysis [17]. But the SVD algorithm is $O(n^3)$ and it would demand huge computational resources or may take a long time to process for large data collections. The complexity of the proposed algorithms will be studied to evaluate the impact on the system performance. The majority of the algorithms that require to process a kernel matrix are training algorithms that can be executed offline. Moreover, training algorithms are not needed to be applied on the complete document collection. That is, a representative sample may be taken from the collection to analyze patterns, structure and relationships, and later the obtained models may be generalized to the whole collection. When possible, parallel or distributed implementations will be considered for algorithms with high complexity.

5. SUMMARY

This paper has presented a research agenda to study and evaluate Multimodal Information Spaces for Content-Based Image Retrieval. The main research question is how can we retrieve visual information from a large multimodal document collection, taking into account that both visual and textual contents may provide useful information to improve the retrieval performance. The use of kernel functions to construct Multimodal Information Spaces is proposed, and a framework based on kernel method solutions will be followed.

Under the proposed framework, different image and text features may be fused in a high-dimensional space, in which a search algorithm may be designed. Each data modality in an image collection will be processed independently and will be integrated using the proposed framework. The image collection to be used is taken from the medical domain in which the multimodal structure may be found in health records and scholarly articles. The evaluation and analysis of standard information retrieval measures is also proposed to assess the contribution of the proposed research.

Bibliography

[1] Stéphane Ayache, Georges Quénot, and Jérôme Gensel. Classifier fusion for svm-based multimedia semantic indexing. pages 494–504. 2007.

[2] A. Bosch, X. Munoz, and R. Marti. Which is the best way to organize/classify images by content? *Image and vision computing*, 25(6):778–791, 2007.

[3] Martin Braschler and Carol Peters. Cross-language evaluation forum: Objectives, results, achievements. *Information Retrieval*, 7(1):7–31, January 2004.

[4] Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. Latent semantic kernels. *Journal of Intelligent Information Systems*, 18(2):127–152, March 2002.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, April 2008.

[6] Xin Fan, Xing Xie, Zhiwei Li, Mingjing Li, and Wei-Ying Ma. Photo-to-search: using multimodal queries to search the web from mobile devices. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 143–150, Hilton, Singapore, 2005. ACM.

[7] Yun Fu, Liangliang Cao, Guodong Guo, and Thomas S. Huang. Multiple feature fusion by subspace learning. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 127–134, New York, NY, USA, 2008. ACM.

[8] Norbert Gövert and Gabriella Kazai. Overview of the initiative for the evaluation of xml retrieval (inex) 2002. In *In Fuhr et al*, volume 6, pages 1–17, 2003.

[9] C. Lau, D. Tjondronegoro, J. Zhang, S. Geva, and Y. Liu. Fusing visual and textual retrieval techniques to effectively search large collections of wikipedia images. pages 345–357. 2007.

[10] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.

[11] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, February 2006.

[12] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[13] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, February 2004.

[14] Rabia Nuray and Fazli Can. Automatic ranking of information retrieval systems using data fusion. *Inf. Process. Manage.*, 42(3):595–614, May 2006.

[15] Trong-Ton Pham, Nicolas Maillot, Joo-Hwee Lim, and Jean-Pierre Chevallet. Latent semantic fusion model for image retrieval and annotation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 444, 439. ACM, 2007.

[16] Lawrence A. Rowe and Ramesh Jain. Acm sigmm retreat report on future directions in multimedia research. *ACM Trans. Multimedia Comput. Commun. Appl.*, 1(1):3–13, February 2005.

[17] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[18] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, New York, NY, USA, 2006. ACM.

[19] Julio Villena-Román, Sara Lana-Serrano, and José C. González-Cristóba. Miracle at imageclefmed 2007: Merging textual and visual strategies to improve medical image retrieval. 2007.

[20] Alexei Vinokourov, David R. Hardoon, and John Shawe-Taylor. Learning the semantics of multimedia content with application to web image retrieval and classification. In *In Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Source Separation*, 2003.

[21] Shengli Wu. Applying statistical principles to data fusion in information retrieval. *Expert Systems with Applications*, In Press, Corrected Proof, 2008.

[22] Tom Yeh, Kristen Grauman, Konrad Tollmar, and Trevor Darrell. A picture is worth a thousand keywords: image-based object search on a mobile platform. In *CHI '05 extended abstracts on Human factors in computing systems*, pages 2025–2028, Portland, OR, USA, 2005. ACM.

Intelligent Web Navigation

Inma Hernández
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Avda. Reina Mercedes s n
41012 Sevilla Spain
inmahernandez@us.es

Abstract

Virtual integration systems retrieve information according to the user's interest. This information is retrieved from several web applications, but it is presented to the user uniformly, in an online process. Therefore, response time is a significant factor. An essential part of any information retrieval system is navigation through pages. Usually web pages contain a high number of links, some of them leading to interesting information, but most of them having other purposes, like advertising or internal site navigation. Traditional crawlers follow every link in each page, in order to analyze the target page, and classify it as interesting or irrelevant. This means having to retrieve, analyze and classify thousands of pages for every single site, which is a costly task. This problem can be solved with the combination of a web page classifier, to distinguish between interesting and irrelevant pages, and a link classifier, which automatically identifies links leading to interesting pages. This kind of navigation is more efficient and has a lower cost than traditional crawlers. Moreover, navigation model is automatically extracted from the site, instead of being handcrafted, reducing the supervision from the user.

Keywords: Navigation, Information Retrieval, Virtual Integration

1. MOTIVATION

Information retrieval obtains all documents relevant to a set of keywords, representing the user's interests, while ignoring all non-related documents. Most IR systems rank the documents according to their relevance to the user, considering features like popularity or correlation to the keywords.

There are two different strategies to locate Web pages: those that are accessible just by following static links are called the *Publicly Indexable Web (PIW)* [32]. However, there are many pages that cannot be accessed this way, instead, they are behind HTML forms, that must be filled in and submitted. These pages constitute a part of the Internet known as the *Deep Web*, and they are our main focus of research.

Automated access to deep web information takes two different approaches, namely *Virtual Integration* (also known as *Metasearch*) and *Surfacing* (also known as *Crawling*) [22]. In virtual integration, users define their interests using keywords that can be as complex as needed, from simple terms to high-level structured queries, and as a response, the system retrieves information related to this keywords. This information is retrieved from many different sources, but it is presented uniformly to the users in a transparent way. This process is online, and therefore response time is an important issue. As opposed to virtual integration, surfacing is an off-line process that intends to collect all pages behind a web form by submitting pre-computed queries, and not taking into account the user's specifical requirements.

Virtual integration systems explore information inside several Web applications, and extract whatever information the user may consider relevant, using information extraction techniques. Therefore, information retrieval and information extraction are two complementary steps in the virtual integration process: the former retrieves all the relevant pages, and the latter extracts required information from these pages.

An essential part of any information retrieval system is the process performed to navigate pages, analyse them, and use the information contained in them to reach further relevant pages. Navigation can be approached in a blind way, e.g., following every possible link in every page. Traditional exhaustive crawlers take this approach, as its goal is to get as many pages as possible.

There is a main drawback to this approach: usually, web pages contain a large number of links, most of them non-relevant to the user (i.e., advertising or links to partner web sites). If the system has to follow all those links, a lot of time will be wasted, making it less efficient.

As a motivating example, Figure 1, shows a web page from a well known e-commerce site. Only three links (the ones marked with a rectangle) lead to relevant pages, while the others have other purposes, like advertising, internal site navigation, or suggestions for the user. As a result, the percentage of useless links is very high. We have observed than most web pages contain more than 70% of links that are useless for information retrieval purposes.

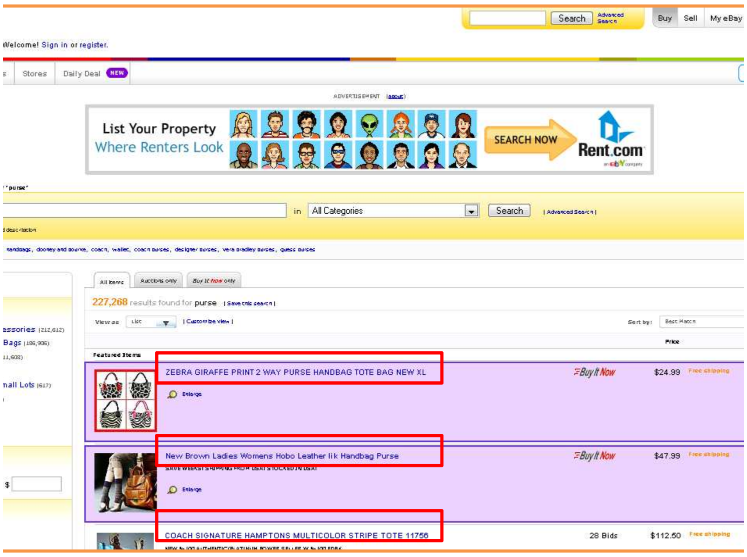


Figure 1: Ebay website.

As opposed to blind navigation, other approaches include some criterion to decide which links to visit and which ones not, therefore reducing the number of irrelevant visited links. Focused crawlers, for instance, hold the criterion of avoiding pages not related to a certain topic (and therefore links leading to them). Relevancy criterions can be handcrafted by the user or automatically decided by the navigator, after analysing the web site, and extracting a navigation model.The latter is what we consider an automated intelligent navigator. In order to decide which pages and links are relevant, intelligent navigators include some reasoning process, usually in the form of a classifier.

The goal of our PhD Thesis is to analyze existing navigation techniques, to select those that are applicable in this context, improve and adapt them if possible, in order to achieve an automated intelligent navigator.

The rest of the article is structured as follows. Section 2 describes related work in the navigation and web page classification areas; Section 3 lists some of the conclusions extracted from the research and concludes the article.

2. RELATED WORK

Figure 2, shows a virtual integration process with intelligent navigation, starting with a set of user keywords which leads to a collection of response pages. Each one of these pages can be a Web page containing the answer to the query made by the user (detail page), or a paginated list of links to detail pages (hub page). In case the server lacks the answer to the query, the response page usually shows an informative message, and optionally some suggestions (no-results page). Finally, when some unexpected error arises, the response page may just contain some error description (error pages). Hence, a classifier is employed to distinguish between the different kinds of response pages (Web Page Classifier), and supports the navigator in discarding both no-results and error pages and retrieving detail pages. If a hub page is detected, it is further analysed in order to discriminate between the links it contains. A usual hub page has different types of links, including advertising, internal site navigation, links leading to detail pages, and links to the previous/following hub page to be recursively analysed (i.e., ”More” or ”Next” links). A link classifier (which we also call Navigator), makes this distinction in order to follow only relevant links. Finally, when detail pages have been identified and retrieved, an information extractor is used to extract relevant structured data.

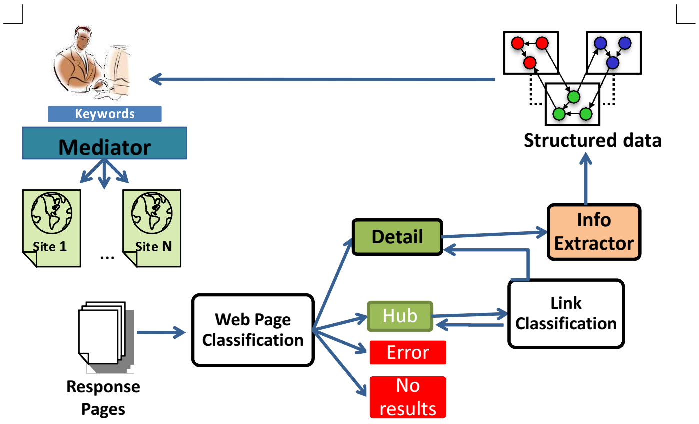


Figure 2: Virtual Integration example.

There are several approaches to the Web classification problem, including content-based classifiers, structure-based classifiers, neighbour-based classifiers, and hybrid techniques. Regarding navigators, there are also some approaches. Some proposals include focused crawlers, recorders, user-defined navigators, and automated navigators. Next, we briefly describe and analyse the related work (summarised in Figure 3).

2.1. Web Page Classification

Web page classification has been extensively researched, and several techniques have been applied with successful experimental results. In general, we classify them according to the type and location of the classification features. There are three main trends in feature types: content-based, structure-based and hybrid classifiers. As for feature location, most approaches obtain features from the page to be classified, while others get them from neighbour pages.

Content-based classifiers ([20], [31] and [34]) categorize a web page according to the words and sentences it contains. This kind of classifiers group all pages within the same topic, assigning them the same class label. As for structure-based classifiers ([3], [6], [10], [16], [19], [33], [36] and [37]), the main feature used to classify pages is their physical organisation of contents, usually expressed in a tree-like data structure, like DOM Tree. Also, there are hybrid approaches, [12] and [23], which take into account both content and structural features. Nevertheless this distinction, there are some abstract classification techniques that sort out pages on the basis of any given feature, e.g., PEBL [39].

All the previous classifiers consider different kinds of features, but in all cases those features are extracted from the page to be classified. There are also classifiers that extract features from the neighbour pages, being the neighbour of a page another page that has a link to the first one. All these proposals are content-based, and usually rely on features such as the link anchor text, the paragraph text surrounding the anchor [15], the headers preceding the anchor, or a combination of these [18]. These techniques are a first approach to the link classification problem.

2.2. Navigation

Crawlers navigate pages by following every link they find. Therefore, the number of pages to be visited grows at a very fast pace. This is useful for certain tasks, like indexing pages for a searching engine, but for our retrieving goal it is not an efficient approach. One example of a traditional crawler is [32].

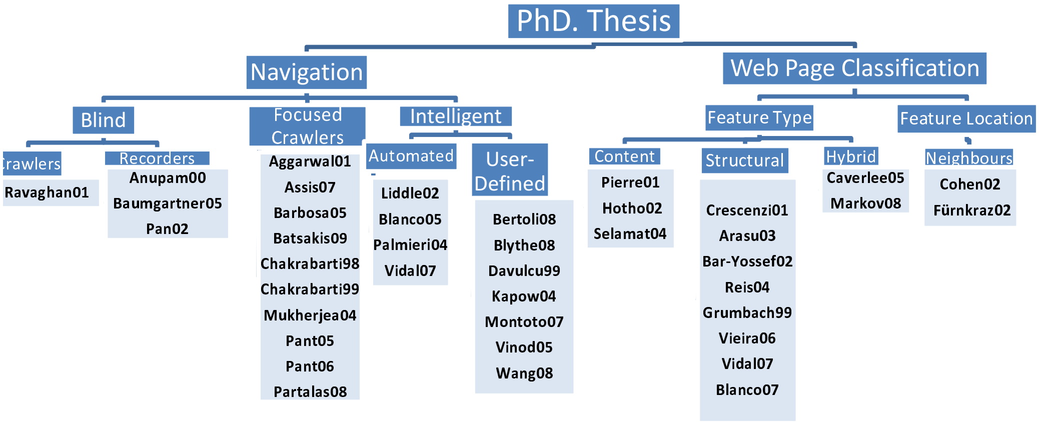


Figure 3: Related Work.

Recorders are also blind navigators, although they are more specific than crawlers, and less flexible. Anupam et. al. [2] present WebVCR, a recording tool with a VCR-like interface to record user's navigation steps through a web site and store them in a file, in order to replay them automatically in the future. Baumgartner et. al. [8] proposes another recording system, based on Lixto, includes an embedded browser in which the user can perform navigation steps easily. Pan et. al. [27] introduced the WARGO system, a wrapper creation tool, and a language to express navigation sequences called NSEQL. Sequences expressed in NSEQL are provided by the user to build a wrapper for information extraction, but the system has a browser-based interface to record the user navigation steps without requiring any programming skills from them. Often, it is necessary to leave some steps undefined until runtime, i.e., when a hub page is shown to the user, the number of clicks the system has to perform on a Next link depends on how many results the search returned. To solve this problem, the WARGO system endows the navigation language with a few extraction capabilities.

The previous proposals interact directly with the web browser interface, so they do not have to deal with problems like scripts, posting forms, having access to pages that require authentication, or navigating sites that keep information about the user, e.g., session IDs. They depend completely upon the user's knowledge, who is responsible for defining the navigation sequences, providing the values to fill in the forms and for redefining the sequences whenever the target web site changes its structure or content. Recording navigation steps makes the navigation inflexible, and not using classifiers results in a more error-prone system.

Focused Crawling is an improvement over traditional crawlers, in the sense that it does include a Web Page Classifier. Focused Crawling ([1], [4], [5], [7], [13], [14], [25], [28], [29] and [30]) combines a traditional crawler with a topical classifier, keeping the focus on pages with topics that have some interest for the user, and therefore reducing the number of useless retrieved pages.

However, focused crawlers improvement is limited to the reduction of useless pages retrieved, but the blind clicking is still an issue. Even when a high number of pages are discarded, they have to be visited and analyzed previously. Moreover, focused crawlers do not classify the functionality of a page, but its topic. When a crawler has submitted a form and gets a response page, both detail and hub pages are the answers to our query, and will probably have the same topic, but they should be processed differently.

User-Defined Navigation is learned by the system in a supervised way, i.e., the user demonstrates how to obtain the interesting information, and the system is able to generalise and acquire this knowledge. EzBuilder [11] is a virtual integration tool used to create information extraction procedures. In order to learn navigation for a web site, the user needs to give an example, submitting forms and navigating to several of the desired pages, and specifying the structure of the data to be extracted. Then, the system builds a model of the user behaviour, that is employed to navigate automatically.

The previous proposal relies on the user, so that the system can learn a navigation model for each site. This model cannot be generalised to other sites automatically and the authors do not report any technique to update it.

Davulcu et. al. [17] presented a technique to express navigation sequences that relies on navigation maps. These are labelled directed graphs in which nodes represent web pages and edges represent actions that can be executed to navigate from one page to another (submitting a form or following a link). The application designer navigates the site, and his or her actions are captured and added to a graph (navigation map). In some cases, the designer needs to provide extended information about the form fields. Some structural changes in the sites can be handled by the system automatically, while other changes makes it necessary to update the maps manually.

Other proposals take a workflow-based approach to represent navigation sequences, although the creation of these sequences is defined by the user. Montoto et. al. [24] model navigation sequences as activities in a work flow diagram. They detect some structural page patterns which frequently occur in web sites. However, they do not use classifiers to automatically distinguish between them, and furthermore the sequence of navigation steps to reach these pages is defined by the user; actually their technique is an extension of [27].

[38] is an example of a semi-automated navigator based on a content classifier. It considers navigation as a sequence of pages, in which one of them is the goal, and the rest are intermediate. Each page belongs to a class, which is defined by a set of keywords, and a set of actions. When the system finds a page belonging to an intermediate class, these actions mark the steps to get to the next page in the sequence. However, if the system finds a page belonging to a goal class, the actions are extraction rules, in order to extract desired information.

Automated navigators are similar to the former proposals, except for the fact that navigation patterns are automatically extracted from web sites, instead of learnt from the user. Liddle et. al. [21] propose a crawling tool, in which forms are submitted, but no specific information is filled in, in order to retrieve as many results as possible. Once the form is submitted, the system is able to automatically distinguish between an indexed list of results, a complete list of results, an informative page with a no-result message, and an error page. When an hub page is detected because of the presence of a Next link, this link is followed iteratively until the last page (or a sufficiently large number of pages is found), and all the pages retrieved are concatenated in a single result page.

This proposal is very simple, since it is only applicable if we wish to retrieve the summarised information contained in hub pages, but not the extended information in detail pages. Furthermore, this tool retrieves all pages, regardless of their relevance to the user's keywords.

For Palmieri et. al., [26], navigation patterns are a graph-based description of the different stages of a user navigation in a web site. They state that most web sites are designed according to the same navigation patterns, e.g., a start page with a link to an advanced search form, which is submitted to obtain a response page. They consider both HTML submission methods, POST and GET, which are expressed as different patterns. Once the form has been submitted, this approach analyses the response page to check if data objects of interest are present, in which case they consider it a hub page. Every hub page has a link leading to the next hub page, until all results are exposed to the user. Some heuristics are presented in order to detect and deal with these links.

However, this approach does not allow for the fact that that hub pages may only contain a brief summary of the data object and a link to a detail page. This proposal limits the possible navigation patterns to a couple of handmade samples, and therefore is not applicable to all web sites.

Vidal et. al. [36], considers that navigation patterns are automatically detected sequences of regular expression URLs that lead to relevant pages. This proposal receives a sample page, which represents the class of pages considered relevant, and it returns a navigation pattern, composed of the sequence of regular expressions that describe URLs that lead to the largest number of relevant pages. Just like the former approach, forms are submitted automatically, and the response page is analysed. If the page belongs to the same class as the one given as example (using a structural classifier), it is considered a relevant page, and it is then retrieved. Small pages are considered to be error pages and they are not processed. Whenever a form is detected, it is submitted again, and process continues recursively. If a large number of links are detected, it is considered a hub page, and a further analysis is needed to select which links will be followed. Links are grouped by their URL similarity, and only one link from each group is followed to check if it leads to a relevant page.

This proposal is semi-supervised, since the process is automated, but the user has to provide a sample detail page, aside from filling in the parameter values for every form submission. This navigator can only reach a small number of possible results, for two reasons: it does not keep all navigation patterns leading to relevant pages, but only the one that leads to the largest number of them, and besides, links leading to other hub pages are not considered, only links to detail pages, so only relevant pages referenced in the first hub page are retrieved. This navigator has to be combined with an iterator in order to retrieve all possible results. Another question to bear in mind is that navigation systems are ad-hoc. If the site changes its URL nomenclature, the system has to be trained again, thus requiring user intervention, to update navigation patterns.

This proposal has some technical problems to be solved, namely, the heuristic for the treatment of form pages can lead to an infinite loop, in cases in which the server includes the original form in every response page. It does not support form submission using the POST method. Finally, this proposal does not support web sites that keep user session information.

[9] is another automated navigator. It is similar to the former: it relies on a structural-based clustering of pages, and it requires a sample relevant page. All pages are grouped in clusters, according to their structure, and links between pages are analysed in order to find relationships between clusters. Whenever a cluster has a high number of outgoing links to the cluster containing the sample, it is considered a hub cluster. This proposal is not designed to crawl pages behind forms, and therefore should be adapted in order to retrieve pages from the deep web.

3. CONCLUSIONS AND RESEARCH QUESTIONS

The main research question posed in this paper is an intelligent navigation system for virtual integration, as opposed to the blind navigation developed by most of the previous works. Our focus is on virtual integration of information within the *Deep Web*. In this context, users specify their interests by means of keywords, and information is retrieved from several sources, regardless of the particular details of each web application. Response pages are analysed, navigating through their links and collecting only those pages that contains the required information.

As we mentioned before, virtual integration is an online process, hence information should be presented to the user in a reasonable response time. Therefore, these systems will indeed find benefits in the application of an intelligent navigator, which avoids visiting useless pages, reducing the cost and improving the efficiency of traditional crawling systems. Summarizing, these are some other research challenges concerning intelligent navigation:

1. Response Web pages have to be classified into relevant and irrelevant, and also into the different roles that they play, which determine how to automatically deal with them.
2. Links in hub pages have to be classified before clicking on them, to identify those that lead to relevant pages. Not only the link anchor text has to be considered, but also the whole context of the link.
3. Link and Web page classifiers in this context are better designed with a lazy execution, that is, classification model is built and updated progressively during the process.
4. It is desirable to develop a navigator with as few interaction from the user as possible. Therefore, learning is unsupervised, or at least, very little supervised. Also, it is interesting to make it as general as possible, not having to build an ad-hoc model for every site, but instead developing a general model that can adapt to most sites just by tuning some parameters. This seems a priori a complex task, however.
5. Advanced post-filtering of relevant detail pages is also a desirable feature. For example, when looking for products in an online store, a user may wish to retrieve only those products whose price lay within a certain range. This means that the navigator needs to be endowed with a lite extraction tool, able to extract from each result at least the necessary data to apply the filter.

6. It is necessary to create a standard data set to evaluate classification and navigation proposals. Most proposals pose some experimental results, measured in terms of precision, recall or some other equivalent metrics. However, very few of them use a well-known data set in their experiments. In most cases researchers collect their own set of pages by issuing queries to a search engine, or using a blind crawler. As the experiments are performed on different sets of pages, the experimental results cannot be compared. There are some well known data sets of already classified pages available for this kind of experiments, e.g., the WebKB collection[1], the Reuters-21578 [2] news collection or the TEL-8 query interfaces collection [3]. [35] reports an attempt to collect a standard data set. Unfortunately, these collections are outdated and updating them is a costly task. We argue that more effort on archiving needs to be done so that new proposals can be compared from an empirical point of view.

Bibliography

[1] Charu C. Aggarwal, Fatima Al-Garawi, Philip S. Yu (2001) *On the design of a learning crawler for topical resource discovery* ACM Trans. Inf. Syst., 19(3):286-309

[2] Vinod Anupam, Juliana Freire, Bharat Kumar, Daniel Lieuwen (2000) *Automating web navigation with the WebVCR.* Computer Networks, 33(1-6):503-517

[3] Arvind Arasu, Hector García-Molina (2003) *Extracting Structured Data from Web Pages.* SIGMOD Conference, 337-348, 2003

[4] Guilherme T. de Assis, Alberto H. F. Laender, Marcos André Gonçalves, Altigran Soares da Silva (2007) *Exploiting Genre in Focused Crawling* SPIRE, 62-73

[5] Luciano Barbosa, Juliana Freire (2005) *Searching for Hidden-Web Databases* WebDB, 1-6

[6] Ziv BarYossef, Sridhar Rajagopalan (2002) *Template Detection via Data Mining and its Applications.* WWW, 580-591

[7] Sotiris Batsakis, Euripides G.M. Petrakis, Evangelos Milios (2009) *Improving the performance of focused web crawlers.* Data & Knowledge Engineering - Elsevier

[8] Robert Baumgartner, Michal Ceresna, Gerald Ledermuller (2005) *Deep Web Navigation in Web Data Extraction* CIMCA/IAWTIC, 698-703

[9] Lorenzo Blanco, Valter Crescenzi, Paolo Merialdo (2005) *Efficiently Locating Collections of Web Pages to Wrap.* WEBIST, 247-254

[10] Lorenzo Blanco, Valter Crescenzi, Paolo Merialdo (2007) *Structure and Semantics of Data-intensiveWeb Pages: An Experimental Study on their Relationships.* J.UCS Special Issue on Wrapping Web Data Islands, 14(11):1877-1892

[11] Jim Blythe, Dipsy Kapoor, Craig A. Knoblock, Kristina Lerman, Steven Minton *Information Integration for the Masses.* J.UCS Special Issue on Wrapping Web Data Islands, 14(11):1811-1837

[12] James Caverlee, Ling Liu (2005) *QA-Pagelet: Data Preparation Techniques for Large-Scale Data Analysis of the Deep Web.* IEEE Trans. Knowl. Data Eng., 17(9):1247-1262, 2005

[13] Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, Jon M. Kleinberg (1998) *Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text.* Computer Networks, 30(1-7):65-74

[14] Soumen Chakrabarti, Martin Van den Berg, Byron Dom (1999) *Focused Crawling: A New Approach to Topic-Specific Resource Discovery.* Computer Networks, 31(11-16):1623-1640

[15] William W. Cohen (2002) *Improving a page classifier with anchor extraction and link analysis.* NIPS, 1481-1488

[16] Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo (2001) *RoadRunner: Towards Automatic Data Extraction from Large Web Sites.* VLDB, 109-118

[17] Hasan Davulcu, Juliana Freire, Michael Kifer, I.V. Ramakrishnan *A layered architecture for Querying Dynamic Web Content.* SIGMOD Conference, 491-502

[18] Johannes Fürnkranz (2002) *Hyperlink Ensembles: A case study in hypertext classification.* Information Fusion, 3(4):299-312

[19] Stéphane Grumbach, Giansalvatore Mecca (1999) *In Search of the Lost Schema.* ICDT, 314-331

[1] WebKB: `http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/`

[2] Reuters: `http://www.daviddlewis.com/resources/testcollections/reuters21578/`

[3] TEL8: `http://metaquerier.cs.uiuc.edu/repository/datasets/tel-8/`

[20] Andreas Hotho, Alexander Maedche, Steffen Staab (2002) *Ontology-based Text Document Clustering.* Künstliche Intelligenz, 16(4):48-54

[21] Stephen W. Liddle, David W. Embley, Del T. Scott, Sai Ho Yau (2002) *Extracting data behind web forms.* ER (Workshops), 402-413

[22] Jayant Madhavan, Loredana Afanasiev, Lyublena Antova, Alon Halevy (2009) *Harnessing the Deep Web: Present and Future.* 4th Biennial Conference on Innovative Data Systems Research (CIDR).

[23] A. Markov, M. Last, A. Kandel (2008) *The Hybrid Representation Model for Web Document Classification.* Int. J. Intell. Syst., 23(6):654-679

[24] Paula Montoto, Alberto Pan, Juan Raposo, José Losada, Fernando Bellas, Victor Carneiro (2007) *A Workflow Language for Web Automation.* J. UCS, 14(11):1838-1856, 2008

[25] Sougata Mukherjea (2004) *Discovering and Analyzing World Wide Web Collections.* Knowl. Inf. Syst., 6(2):230-241

[26] Juliano Palmieri Lage, Altigran S. da Silva, Paulo B. Golgher, Alberto H.F. Laender (2004) *Automatic generation of agents for collecting hidden Web pages for data extraction.* Data Knowl. Eng., 49(2):177-196

[27] Alberto Pan, Juan Raposo, Manuel Álvarez, Justo Hidalgo, Ángel Viña (2002) *Semi-Automatic Wrapper Generation for Commercial Web Sources.* Engineering Information Systems in the Internet Context, 265-283

[28] Gautam Pant, Padmini Srinivasan (2005) *Learning to crawl: Comparing classification schemes* ACM Trans. Inf. Syst., 23(4):430-462

[29] Gautam Pant, Padmini Srinivasan (2006) *Link Contexts in Classifier-Guided Topical Crawlers.* IEEE Trans. Knowl. Data Eng., 18(1):107-122

[30] Ioannis Partalas, Georgios Paliouras, Ioannis P. Vlahavas (2008) *Reinforcement Learning with Classifier Selection for Focused Crawling* ECAI, 759-760

[31] John M. Pierre (2001) *On the Automated Classification of Web Sites.* CoRR, cs.IR/0102002

[32] Sriram Raghavan, Hector Garcia-Molina (2001) *Crawling the hidden web.* VLDB, 129-138

[33] Davi de Castro Reis, Paulo B. Golgher, Altigran S. da Silva, Alberto H. F. Laender (2004) *Automatic Web News Extraction Using Tree Edit Distance.* WWW, 502-511

[34] Ali Selamat, Sigeru Omatu (2004) *Web page feature selection and classification using neural networks.* Inf. Sci., 15869-88

[35] Mark P. Sinka, David W. Corne (2002) *A Large Benchmark Dataset for Web document clustering.* Soft Computing Systems: Design, Management and Applications, 87 (2002)

[36] Márcio Vidal, Altigran S. da Silva, Edleno S. de Moura, Joao M.B. Cavalcanti (2007) *Structure-Based Crawling in the Hidden Web.* J. UCS, 14(11):1857-1876

[37] Karane Vieira, Altigran S. da Silva, Nick Pinto, Edleno S. de Moura, Joao M. B. Cavalcanti, Juliana Freire (2006) *A Fast and Robust Method for Web Page Template Detection and Removal.* CIKM, 258-267

[38] Yang Wang, Thomas Hornung (2008) *Deep Web Navigation by Example.* BIS (Workshops), 131-140

[39] Hwanjo Yu, Jiawei Han, Kevin Chen-Chuan Chang (2004) *PEBL: Web Page Classification without Negative Examples.* IEEE Trans. Knowl. Data Eng., 16(1):70-81

Ontology-based terminology management for transitive translations focusing on NEs

Fumiko Kano Glückstad
Department of International Language Studies and Computational Linguistics
Copenhagen Business School
Dalgas Have 15, DK-2000 Frederiksberg, Denmark
fkg.isv@cbs.dk

I demonstrate that there are two types of transitive translations of Named Entities (NEs), both of which should be handled in the process of Cross Lingual Information Retrieval (CLIR). An official transitive translation is defined as a translation made through an official English translation often provided from an authorized local entity. A lexical translation is a direct lexical translation from a source language to a target language. However, a lexical translation also requires a transitive translation when a language pair is a rare combination having inadequate language resources. Hence I define it as a lexical transitive translation. I assess the inconsistency level of the official- and lexical transitive translations of NEs and propose an ontology-based CLIR solution referred to as *triangulated terminology management*.

Multilingual information retrieval, CLIR, transitive translation, pivot translation, named entity disambiguation

1. BACKGROUND

My research issue has been raised by the question: Is it possible to identify local first-hand information produced in non-English speaking countries from Japanese queries translated from their official English information sources? Specifically, the issue is rooted in a plurality of inconsistencies found between Japanese translations made through the direct lexical translation from Danish to Japanese and Japanese translations made through the transitive translation using official English translations as source. A typical example of such a translation problem is illustrated where the formal English name of the Danish authority “Økonomistyrelsen” is “The Danish Agency for Governmental Management.” The Danish originated name, “Økonomistyrelsen”, will most likely be translated into a completely different Japanese expression through lexical English translations, “Economy Agency (*keizai-tyou*)” using available language resources such as Danish-English and English-Japanese dictionaries. Eventually, it becomes increasingly difficult for Japanese readers to identify the original Danish NE in the process of CLIR due to inconsistent Japanese translations. Hence my research addresses the following three issues: 1) evaluation of the inconsistency level between the direct Japanese translation of Danish NEs and of their officially translated English; 2) developing an ontology-based solution to identify an original entity from inconsistent translations based on a triangulated terminology management approach; and 3) eventually and hopefully, to identify a base-line CLIR system that can integrate the triangulated terminology management approach. This paper is only addressing the initial phases raised by issue 1.

2. THE OFFICIAL TRANSITIVE TRANSLATION AND LEXICAL TRANSITIVE TRANSLATION

In CLIR, one of the basic methods in query translation is called dictionary-based translation. The problem with this method is that there are inherently insufficient language resources available for most language pairs that are part of rare combinations. Hence, it is required to employ a transitive translation technique using a so-called pivot language. Gollins and Sanderson [1] pointed out that, since a transitive translation in CLIR is based on a simple word-by-word translation approach, it increases the likelihood of translation errors, caused mainly by incorrect identification of the sense of ambiguous words. Ballesteros [2] examined the impact of transitive translations and discovered that using simple word-by-word transitive translations from Spanish to French via English degraded performance by 91% when compared to a direct bilingual translation from Spanish to French. Gollins and Sanderson [1] introduced an approach to reducing errors by combining translations from two different transitive routes, a process known as *lexical triangulation*. Their results showed that the lexical triangulation approach to the transitive translation eliminated the differences in retrieval between transitive translated queries and equivalent direct translated queries.

However, considering the aforementioned specific example of the Danish NE, “Økonomistyrelsen”, there are two types of transitive translations and the solution proposed by Gollins and Sanderson [1] only addresses issues arising from the lexical translation from Danish to Japanese. It means that it is necessary to clearly distinguish the transitive translation using an official English translation as inter-lingua from the transitive translation based on a lexical translation. Hence in my research, I define the transitive translation using an official English translation as an *official transitive translation* and the transitive translation based on a lexical translation as a *lexical transitive translation*. In this work, I report the preliminary survey of measuring frequency and semantic similarity of the official- and the lexical transitive translations of Danish NEs.

In order to identify inconsistencies between the *official*- and *lexical transitive translation* of original Danish NEs, I compared differences between official English translations and lexical English translations of names of Danish governmental organizations (i.e. ministries and institutions under the ministries), most of which provide official English names of their organizations. For performing a *lexical translation* of original Danish NEs into English, I used one of the most popular Danish-English dictionary series entitled "Gyldendals Røde Ordbøger". Regarding the *lexical translation*, I defined the following rules: 1) NEs consisting of several words should be translated word-by-word; 2) If the dictionaries propose an English translation equal to the corresponding official English translation, the official English expression should be applied. Accordingly, I translated all of 70 selected Danish NEs into English and extracted 26 English *lexical translations* that were not identical to their respective *official translations*. Since these English translations of NEs are Multi-Word Expressions, I further decomposed them into each lexical unit (word) and enlisted the inconsistent word pairs that were scope for further analysis. For comparing the semantic similarity of these word pairs, I used the basic Path Length measure provided on the web interface of the WordNet::Similarity [3]. The results showed the semantic distance in most of the word pairs produced via official- and lexical English translations. For example the official English expression of "Ministeriet for Videnskab, Teknologi og Udvikling (Ministry for Science, Technology and Development)" is "Ministry of Science, Technology and Innovation". In the same way, "Forsknings- og Innovationsstyrelsen (Research and Innovation Agency)" is "Agency for Science, Technology and Innovation". The semantic distances of word pairs "innovation vs. development" and "research vs. science" are respectively shown in FIGURE1.

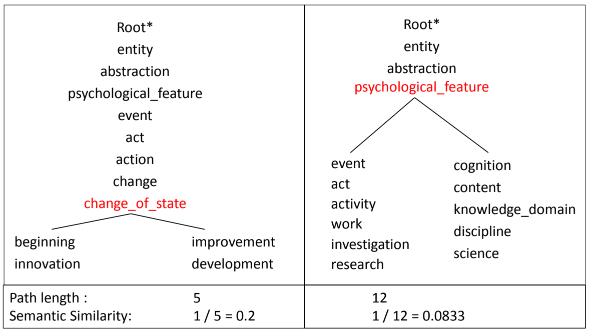


FIGURE 1: Examples of Semantic similarity

4. OUTLOOK

My study shows that the similarity measures based on Path Length indicate the degree of inconsistency level between English translations made through a so-called *official translation* and a so-called *lexical translation*. The next noteworthy question is how a Japanese translation of these pairs of English translations will turn out. My initial assumption is that these Japanese translations will create expressions with an even deeper level of inconsistency (i.e. FIGURE 2). It means that it will be increasingly difficult to identify the original Danish NEs from various Japanese translations. If there were universal rules defining "a name should always be translated based on the lexical meaning of its original language", these inconsistencies would potentially be tremendously reduced. However, the reality is unfortunately far from that. Usually, the decision of names and their translations involves a plurality of issues, such as political (domestically, internationally), cultural, social and so on. It means that problems originating from both *official*- and *lexical transitive translations* should be carefully dealt with in terms of a so-called Named Entity Disambiguation.

As a solution, I propose an ontology-based triangulated terminology management approach. The approach is based on the idea that a country specific NE has a unique ontological structure, since a named entity is per definition unambiguously defined on a global scale. For example, the Danish governmental organizations are existing according to a Danish governmental structure that is uniquely defined in this country. It means that the ontological structure is unique even though each named entity is expressed in different languages. Therefore, an ontology-based terminology database consists of three layers: a) each NE expressed in a source language, b) its official expression in an inter-lingual language (usually in English), and c) all possible expressions in a target language (FIGURE 3). Each entity in an ontological hierarchy should contain metadata specifying country, timeframe, structural relation etc. These three layers should have a triangulated relationship as shown in FIGURE 4. The key issue is that the name of an entity expressed in a source language and an official expression in an inter-lingual language should have a relationship linking them like "is translation of" each other. However, an expression in a target language that "is translation of" either a name of an entity expressed in a source language or an official expression in an inter-lingual language is uni-directionally linked and hence cannot be traced the other way around. A frame for expressions in a target language should contain all possible translations from any available corpora in the target language. It is my aim to establish a triangulated terminology database in the Danish e-government domain based on an ontology-based terminology management system developed by Copenhagen Business School [4]. As the next step, I would like to investigate and to identify a base-line CLIR system that can integrate the triangulated terminology management approach.

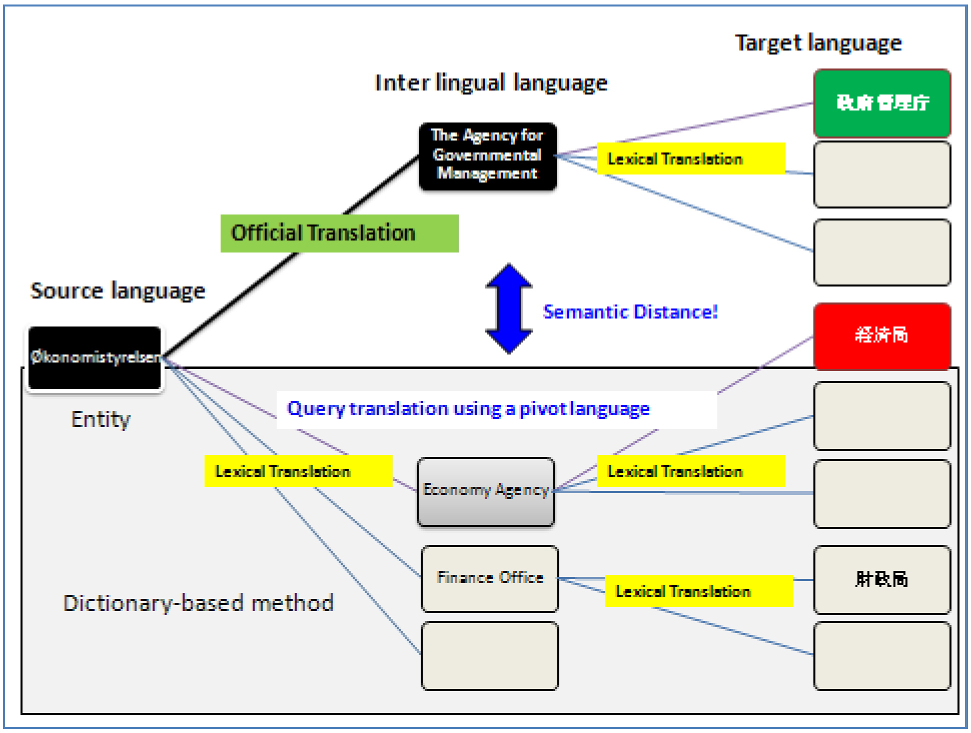
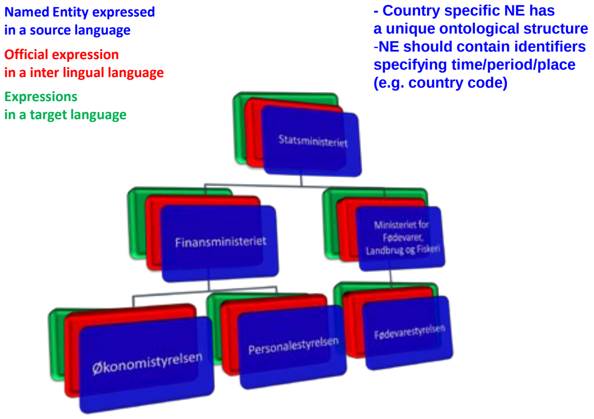


FIGURE 2: Inconsistent pivot- and transitive translations



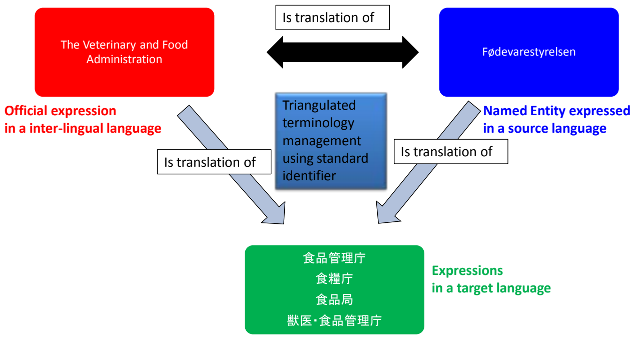


FIGURE 3: Ontology-based terminology management

FIGURE 4: Triangulated terminology management

REFERENCES.

[1] Gollins, T. and Sanderson, M. (2001) Improving Cross Language Information Retrieval with Triangulated Translation, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, United States:pp.90-95

[2] Ballesteros, L. (2001) Cross Language Retrieval via transitive translation, In Croft W. B. (ed). *Advances in Information Retrieval: recent Research from the CIIR*, Kluwer Academic Publishers, pp.203-234

[3] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004) WordNet::Similarity – Measuring the Relatedness of Concepts. Available from: http://search.cpan.org/dist/WordNet-Similarity.

[4] Madsen, B, Thomsen, H. and Wenzel, A (2006) i-Term for NORDTERM *5th International Conference on Language Resources and Evaluation (LREC 2006)*, Workshops Proceedings: W16 Terminology Design: Quality Criteria and Evaluation Methods (TermEval). Genova, Italy

Classifying High Quality Photographs by Creative Exposure Themes

Supheakmungkol SARIN and Wataru KAMEYAMA
Graduate School of Global Information and Telecommunication Studies, Waseda University
mungkol@fuji.waseda.jp, wataru@waseda.jp

Abstract

In this paper, we propose to utilize contextual camera setting parameters at the time of capture to perform the classification task of high quality photographs. With supervised machine learning algorithm, we build a model that can classify high quality photographs into six creative exposure themes which are commonly known and used by the professional photographers. Our experiments give us an encouraging result.

Keywords: Aesthetics, Experimentation, Classification

1. INTRODUCTION

In this age of the digital photograph explosion, media companies - especially stock photo agents, advertising and printing companies - have huge collections of high quality photographs. The task of selecting a suitable picture for a targeted theme is, and will still be, a burden, even though there are annotations in the collection. For instance, how does one select an image that depicts *freezing action*, an image that has a *great depth of field* or an image that *implies motion* for a front cover of a magazine? To lessen this difficulty, we are looking at the problem of classification from the professional photographer's perspective.

2. CONSIDERATIONS

2.1. Exposure and Patterns in High Quality Photographs

In photography, the exposure control being a process of controlling light projecting to camera's digital sensor is the main actor to successful photography. Exposure is determined by three settings - shutter speed, lens aperture and ISO. Correct combination of these three will result in a good photo - a well exposed photo. Obviously, there are many of such combinations that can result in a well exposed photo. However, among them only a few can give interesting photographs. In his book entitled *Understanding Exposure* [1], Peterson distinguishes seven classes of high quality photographs by exposure theme. He calls them *creative exposure themes*. Furthermore, he discusses the characteristics and the rules that can be used to produce those images. In this study, we focus only on six exposure themes because we have limited number of photos that correspond to the seventh theme in our dataset. The following explains each theme and Figure 1 shows the example images of those themes.

- *Story Telling (ST)*: when we want great depth of field with all objects inside to be neat and clear. It is usually done using wide angle lens and small aperture.
- *Who Cares (WC)*: when the depth of field is not a concern and when subjects are at the same distance from the lens. It is usually done with middle range aperture.
- *Isolation or Single Theme (I)*: when we want to focus on a specific subject. It is usually done with a large aperture open. Usually, the unfocused part is blur.
- *Freeze action (FA)*: when we want to freeze and capture the moment. This is usually done using very fast shutter speed.
- *Imply motion (IM)*: when we want to convey motion to the audiences. This is usually done using very slow shutter speed.
- *Macro or Close-up (M/C)*: when we want the great detail of the subject or just part of it in close proximity. Usually, we want to record the image from 1/10 to 10 times or more of the actual size. The image often lacks of depth of field.

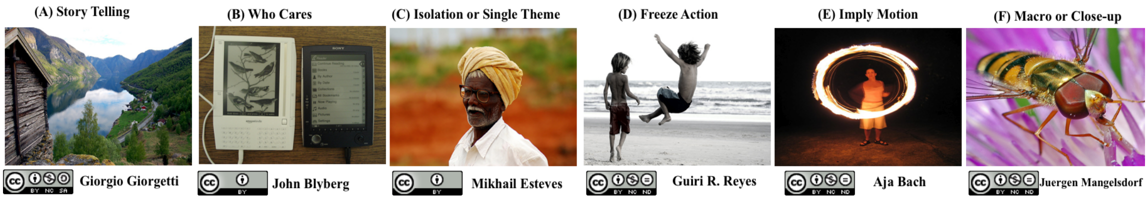


FIGURE 1: Example images of the six creative exposure themes

2.2. Camera Setting Parameters

As described earlier, lens aperture, shutter speed, and ISO play important roles in creating a correct exposure for each theme. Fortunately, unlike conventional camera, current modern digital cameras are equipped with many sensors. Many kinds of information are recorded at the same time when a photograph is taken. If we make an analogy of those sensors to our human eyes, this captured information represents the *intention* of the (professional) photographers. Usually, when taking a photo, photographer has in mind which type of photo he or she is going to make and configure the camera setting accordingly. Specifically, two main things can be extracted: *photographer's intent* and the *condition in which the image is captured*. EXIF specification [2], which is universally supported by most of digital cameras, enables these settings. Some of the important parameters which professional photographers usually refer to and which can be found in the EXIF header of the each image file are: *Lens Aperture*, *ISO*, *Exposure Time/Shutter Speed*, *Date and Time*, *Focal Length*, *Metering Mode*, *Camera Model*, *Exposure Program*, *Maximum Lens Aperture*, *Exposure Bias*, *Flash*, etc.

3. METHODOLOGY, IMPLEMENTATION AND RESULTS

With the above considerations, there is an obvious relationship between creative exposure themes and some of the camera setting parameters. Thus, in this work, we propose to categorize the photographs into six creative exposure themes and tackle the problem computationally and experimentally using statistical learning approach by applying on the camera setting parameters.

3.1. Dataset and Extracted Features

We use the recent MIR Flickr 25000 test collection [3]. The photos in the collection are selectively taken from Flickr[1] based on their high interestingness rate. As a result the image collection is representative for the domain of original and high quality photography. 75% of them have the 5 major settings namely, *Aperture*, *Exposure Time*, *Focal Length*, *ISO Speed* and *Flash*. We use all of these features in this work. Based on the camera model found in EXIF, we also distinguish Point-and-Shoot cameras with Digital Single Lens Reflection ones. For our study, a subset of the collection (2736 photos) is labeled into the six themes. The labeling process is done manually based on the strong correspondence of the visual expression of each of the photos to the six creative exposure themes. One problem that we faced during the labeling process is that some photos can be attributed to multiple themes. For that we put the photo to the most suitable class.

3.2. Model Building, Evaluation and Results

We divide our dataset into training (2/3) and testing sets (1/3). We carefully create the random splits within each class so that the overall class distribution is preserved as much as possible. With the training set, several machine learning algorithms such as Decision Tree, Forest, SVM and Linear combination were used to train the dataset and create the models automatically. Finally, to evaluate the models, we test them with the testing set. The confusion matrix is computed. We calculate the performance of each established model by the following measures: precision as percentage of positive predictions that are correct, recall/sensitivity as percentage of positive labeled instances that were predicted as positive, specificity as percentage of negative labeled instances that were predicted as negative, and accuracy as percentage of predictions that are correct. Decision Tree which is rather simpler than other models gives the best performance of all. Due to limited space, we show only our best result. Figure 2 depicts our generated model while Table 1 and Table 2 show the performance of the model.

[1] Flickr webiste: http://www.flickr.com

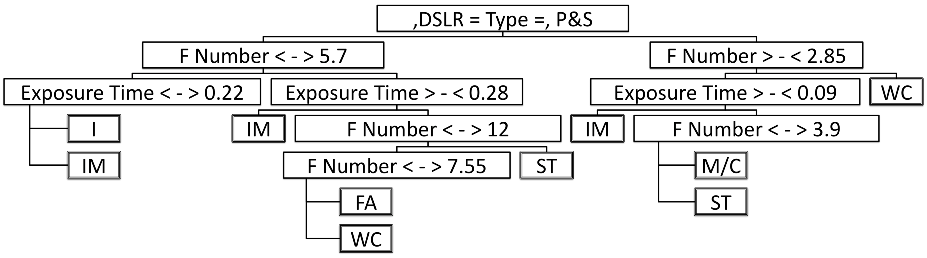


FIGURE 2: Generated Decision Tree Model

TABLE 1: Confusion Matrix

		Actual Themes					
		FA	I	IM	M/C	ST	WC
Predicted Themes	FA	15	0	1	5	0	0
	I	74	169	10	41	1	1
	IM	0	0	58	4	6	2
	M/C	6	0	2	7	0	0
	ST	30	0	5	23	127	0
	WC	28	0	13	30	1	253

Even though we used only the EXIF parameters and the camera type in this study, we obtained an encouraging result. We also observed that the model generated by Decision Tree only use 3 parameters namely, F Number, Exposure Time and Camera Type. Also, the generated model corresponds to what describes by Peterson. This raises question whether more features which might demand high computational costs are needed.

4. CONCLUSION

We present a technique to classify high quality photos from professional photographer's perspective without using any conventional low-level features. Recently, there have been research efforts in using EXIF metadata to classify and annotate photographs. Ku et al. used scene modes for image scene classification [4]. Sigha et al. utilized optical information for the task of photo classification and annotation [5]. However, to the best of our knowledge, this work on classification of high quality photographs by focusing on the creative exposure themes is the first attempt so far.

Though we obtained a reasonable performance using very few features, for our future work, we would like to see how the integration with other type of features could help this task even more with regards to the trade-off of computational costs. For our immediate study, content-based features such as color, texture, shape, and scene description will be integrated. We also would like to perform our experiment on larger dataset with multiple annotators to avoid any bias.

REFERENCES

[1] B. Peterson. Understanding Exposure [Revised Edition]. AMPHOTO Book, 2004

[2] EXIF Specification: http://www.exif.org/specifications.html [Accessed on June 23, 2009]

[3] M. J. Huiskes, M. S. Lew. The MIR Flickr Retrieval Evaluation. In Proc. ACM MIR, 2008

[4] W. Ku, M. S. Kankanhalli and J. Lim. Using Camera Settings Templates ("Scene Modes") for Image Scene Classification of Photographs Taken On Manual/Expert Settings, Proceedings of 8th Pacific-Rim Conference on Multimedia, LNCS 4810, pp. 10 - 17, 2007.

[5] P. Sinha and R. Jain. Classification and Annotation of Digital Photos using Optical Context Data. ACM International Conference on Content-Based Image and Video Retrieval, pp. 309-318, Niagara Falls, Canada.

TABLE 2: Precision, Recall/Sensitivity, Specificity and Accuracy Rate (Let $TP : TruePositive, TN : TrueNegative, FP : FalsePositive, FN : FalseNegative$)

	FA	I	IM	M/C	ST	WC	Average
$Precision = \frac{TP}{TP+FP}$	0.71	0.57	0.82	0.46	0.68	0.77	0.67
$Recall = \frac{TP}{TP+FN}$	0.09	1	0.65	0.063	0.94	0.98	0.62
$Specificity = \frac{TN}{TN+FP}$	0.99	0.82	0.98	0.99	0.92	0.89	0.93
$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	0.84	0.86	0.95	0.87	0.92	0.91	0.89

Similarity Learning in Nearest Neighbor and Application to Information Retrieval

Ali Mustafa Qamar and Eric Gaussier
LIG (Laboratoire d'Informatique de Grenoble)
Université Joseph Fourier
ali-mustafa.qamar@imag.fr, eric.gaussier@imag.fr

Abstract

Many people have tried to learn Mahanalobis distance metric in kNN classification by considering the geometry of the space containing examples. However, similarity may have an edge specially while dealing with text e.g. Information Retrieval. We have proposed an online algorithm, *SiLA* (Similarity learning algorithm) where the aim is to learn a similarity metric (e.g. cosine measure, Dice and Jaccard coefficients) and its variation *eSiLA* where we project the matrix learnt onto the cone of positive, semi-definite matrices. Two incremental algorithms have been developed; one based on standard kNN rule while the other one is its symmetric version. *SiLA* can be used in Information Retrieval where the performance can be improved by using user feedback.

Keywords: Similarity learning, kNN, Information Retrieval, Machine Learning

1. INTRODUCTION

Many works have tried to improve the kNN algorithm by considering the geometry of the space containing examples. Most of these works learn Mahanalobis distance metric, a variation of Euclidean distance. The Mahanalobis distance between two objects x and y is given by:

$$d_A(x, y) = \sqrt{(x-y)^T M (x-y)}$$

However, similarity should be preferred over distance in many practical situations, e.g. text classification, information retrieval as was proved by our results on different datasets [4].

2. PROBLEM FORMULATION

The aim here, is to learn a similarity metric for kNN algorithm. Let x and y be two examples in $\mathbb{R}^p$. We consider similarity functions of the form:

$$\mathsf{s}_A(x, y) = \frac{x^T A y}{\mathsf{N}(x, y)} \tag{1}$$

where A is a $(p \times p)$ matrix (symmetric or asymmetric) and $\mathsf{N}(x, y)$ is a normalization which depends on x and y. Equation 1 generalizes several different similarity functions (cosine measure (by replacing matrix A with the identity one), Dice coefficient, Jaccard coefficient)

3. *SILA* (SIMILARITY LEARNING ALGORITHM) AND *ESILA*

SiLA is based on voted perceptron developed by [3] and used by [2]. Figure 1 illustrates the notion of separability we are considering. In 1(a), the input point is separated, with $k = 3$, whereas it is not in 1(b) as it is closer both to points from the class it belongs as well as differently labeled examples. The separation does not need to take place in the original input space, but rather on the space induced by the metric defined by A. 1(c) illustrates what we are aiming at: moving the target points closer to the input point, while pushing away differently labeled examples.

When an input example is not separated from examples belonging to different classes, the current A matrix is updated by the difference between the coordinates of the target neighbors and the

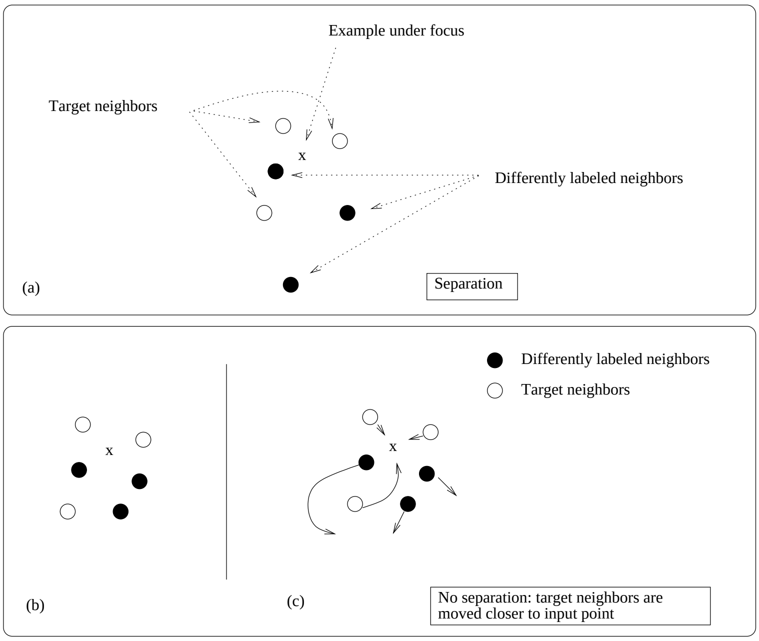

Example under focus
Target neighbors
x
Differently labeled neighbors
(a)
Separation
Differently labeled neighbors
Target neighbors
x
x
(b)
(c)
No separation: target neighbors are moved closer to input point

FIGURE 1: In (a) the input point is separated with $k = 3$, whereas it is not in (b). (c) illustrates the process we aim at: moving target points closer to the input point, while pushing away differently labeled examples.

closely differently labeled examples. This update corresponds to the standard perceptron update. However, in case of correct classification, the weight corresponding to the current A matrix is increased by 1. The weighted matrices are then used to classify unseen test examples. Two prediction rules have been developed: one is based on standard kNN (kNN-A) while the other one considers same number of examples in different classes (SkNN-A). It has been proved that the no of mistakes are bounded and the algorithms can generalize well beyond training examples.

In *eSiLA*, A matrix is orthogonally projected on the cone of positive semi-definite matrices inspired from *POLA* [5]. This projection guaranties convergence and generalization of the algorithm. However, *eSiLA* is similar to *SiLA* in all other aspects.

4. EXPERIMENTAL VALIDATION

SiLA and *eSiLA* were tested on eight standard test collections, namely *Balance*, *Wine*, *Iris*, *Ionosphere*, *Soybean*, *Glass*, *Pima* and *20-Newsgroups*, where first seven were obtained from UCI [1]. 5-fold nested cross validation was used to learn the matrix A for the UCI datasets owing to their small size. We used two prediction rules for the experiments. In the first one, the classification is based on the k nearest neighbors (*kNN* rule) while the second one (*SkNN*) is based on the difference of similarity between k nearest neighbors from the same class and k from differently labeled classes. The results, given in table 1 demonstrate that similarity should be preferred over distance on non-textual collections also like *Balance* (gain of 7.6%), *Wine* (gain of 8%), *Iris* (gain of 0.9%), *Ionosphere* (gain of 1.7%) etc.

The results further show that the algorithm *eSiLA* performs better as compared to standard kNN on *Wine* (gain of 1.9% with *SkNN-A*), *Ionosphere* (gain of 1.9% with both *kNN-A* and *SkNN-A*) and *Pima* (gain of 0.8% with *SkNN-A*) . All methods have comparative performance on *Soybean* and *Glass* (since base accuracy is already too high with just using cosine). *SiLA* improved the base results (with *kNN-cos*) on *Balance*, *Wine*, *Ionosphere* and *News*.

eSiLA performs better than *SiLA* on *Wine* (gain of 1.2%), while they are comparable on *Ionosphere*, *Pima* and *Soybean*.

TABLE 1: Results on all collections

	kNN-eucl	kNN-cos	kNN-A	SkNN-cos	SkNN-A
Balance	0.883	0.959	0.979	0.969	**0.983**
Wine	0.825	0.905	**0.916**	0.909	**0.916**
Wine (eSiLA)	0.825	0.905	**0.926**	0.909	**0.928**
Iris	0.978	**0.987**	**0.987**	0.982	**0.987**
Ionosphere	0.854	0.871	**0.911**	0.871	**0.911**
Ionosphere (eSiLA)	0.854	0.871	**0.914**	0.871	**0.914**
Soybean	**1.0**	**1.0**	0.994	0.989	0.989
Soybean (eSiLA)	**1.0**	**1.0**	**1.0**	0.989	0.989
Glass	**0.998**	**0.998**	0.997	**0.998**	0.997
Pima	**0.698**	0.652	0.647	0.665	**0.678**
Pima (eSiLA)	**0.698**	0.652	0.659	0.665	**0.673**

5. SIMILARITY LEARNING AND INFORMATION RETRIEVAL

SiLA or *eSiLA* can be used in Information Retrieval, where the matrices can be tuned by incorporating user feedback. The similarity is calculated between a query q and a document d. The basic theme rests the same: try to bring target documents (documents relevant to q) closer to q while pushing away irrelevant documents which in turn yields matrix A. The top rated documents are presented to the user who can then change the order. This order is learnt by updating the weights in the same way as in *SiLA* and *eSiLA*.

6. CONCLUSION

In this paper, we have discussed an approach to apply similarity learning algorithms namely, *SiLA* and *eSiLA*, in the context of Information Retrieval. Both *SiLA* and *eSiLA* not only outperform standard euclidean distance on some collections, but also improve the base results using standard cosine. We have described a mechanism to incorporate user feedback in order to update the similarity matrix.

REFERENCES

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[2] M. Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 1–8, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[3] Y. Freund and R. E. Schapire. Large margin classification using the perceptron algorithm. *Mach. Learn.*, 37(3):277–296, 1999.

[4] A. M. Qamar, É. Gaussier, J.-P. Chevallet, and J.-H. Lim. Similarity learning for nearest neighbor classification. In *ICDM*, pages 983–988, 2008.

[5] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng. Online and batch learning of pseudo-metrics. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, New York, NY, USA, 2004. ACM.

Summarization as a Means of Information Access: Utilizing Semantic Metadata

Eugenie Giesbrecht
FZI Research Center for Information Technology
Haid-und-Neu-Str. 10-14, 76131, Karlsruhe, Germany
www.fzi.de/ipe
giesbrecht@fzi.de

Abstract

The existing search engine interaction paradigm of typing in keywords and getting an enormous list of links is not suited for a lot of information seeking tasks. We see synthesis or summarization of information to satisfy users' information needs as an important step on the way to next generation information access systems. The idea is to explore the alternative geometrical models of meaning [14] based on the theory of Quantum Mechanics and Hilbert Spaces as a unifying framework for integrating semantic metadata into retrieval and summarization.

Keywords: Information Retrieval, Summarization, Question Answering, Geometrical Models of Meaning

1. MOTIVATION

A well-known paradigm of querying documents in a classical information retrieval system is by inputting keywords and matching them against the terms by which the documents are indexed. In reply, the user receives a list of links to be consulted. This is rather similar to searching, borrowing and looking for relevant information in books in the library [10]. Paradoxically, ”information” retrieval has established itself as a pure document retrieval.

The so-called information overload causes the traditional library paradigm of information retrieval systems to be reconsidered. As discussed in [13, 10], the existing search engine interaction paradigm of typing in keywords and getting a list of results is not suited for a lot of information seeking tasks. Though in the meantime we are able to extract named entities and patterns in the textual information thanks to text analysis research, this kind of semantic metadata is still insufficiently integrated in information seeking technology. Having detected certain entities and relations in the text is not an ultimate goal. The next challenge is to make use of this information to satisfy the user's information needs. The vision is that information access systems should more directly answer our information needs by information extracted from the documents and as far as possible processed and synthesized into a coherent answer [10].

Thereby, synthesis or summarization of information as an answer to a user's query is an important step on the way to *cooperative* information access systems.

2. BACKGROUND AND RELATED WORK

This work is on the intersection of at least three distinct communities, where the research has been conducted more or less independently of each other - information retrieval, open domain question answering and summarization - and will draw, amongst others, upon the research from text mining and cognitive information processing. As a representational formalism, multi-dimensional geometrical models of meaning are being explored.

In the following, we briefly mention relevant work in corresponding disciplines and figure out the weaknesses and potential intersections of current approaches. Section 3 points at some of the research questions that emerge out of the latter.

2.1. Information Retrieval

At the Demo'08 panel session[1] discussing the future of the web, it has been emphasized by the representatives of the biggest search companies[2] that the search should become "task-centric" or "wish fulfilling". "The search engine should 'read' and synthesize the information to solve the intent", said Prabhakar Raghavan. The latter implies a fundamental change in the design assumptions of search systems by moving them from document search to the tasks for which people employ search. This is the idea behind this research - to synthesize information, presumably in a structured way, depending on the user's current intention.

2.2. Summarization and Open Domain Question Answering

Summarization is a rather new paradigm in information access and retrieval research. At the same time there is a long term tradition of research on summarization in NLP community, also specifically for purposes of open domain question answering. Automatic summarization here is a task of extracting the most important content from information sources and presenting it to the user in a condensed form and in a manner sensitive to the user's or application's needs [9].

The research in text summarization goes back to 1958, motivated by work of Luhn [8] who developed the first system of sentence extraction and building extractive summaries. The early work in open domain question answering goes back to the 90s. The ultimate goal of the latter has been to build systems that are able to answer any question in any domain. However, similarly to summarization and due to the inherent complexity of the task, this research focused mostly on extracting passages or sentences that have been ranked as most relevant to the question. The challenge here ended up in trying to get the most relevant sentence or passage as first-ranked. Consequently, the dominant approach in both summarization and question answering is still extraction rather then real abstraction, though both communities realize the need for abstraction and true synthesis of information [6, 11].

2.3. Dual Document Representations

The idea to move away from the *bag-of-words* in IR research, e.g. by mapping terms to concepts or accessing documents by extracted pieces of information, has been in the air for a while. Harris [5] already in 1959 proposed to extract certain relations from scientific articles by means of NLP and to use them for information finding. In order to achieve a kind of "conceptual" search, indexing strategies where the documents are indexed by concepts of WordNet [4], of Wikipedia [3] or ontology [1] have been used. It is just that the time is ripe now and information extraction technology has matured enough to use it on a large scale [10]. Most of these approaches have used either the one or the other way of "homogeneous" document representation. There are some attempts on the way to realize dual document representations by utilizing statistical language modelling (e.g. [2]). We are not aware, however, of any work aiming at integratation of dual text representations by means of Hilbert Spaces.

3. RESEARCH DIRECTIONS

The goal of this thesis is to explore the ways to use the explicit semantic information, i.e. semantic metadata attached to the documents, not only to improve retrieval, but also to propose new ways of answering users' complex information needs in an "information overload" era in a cooperative way, i.e. by summarizing and allowing exploration based on the user's information need.

[1] http://www.demo.com/watchlisten/videolibrary.html?bcpid=1127798146bclid=1782597597bctid=1790936412

[2] Peter Norvig from Google and Prabhakar Raghavan from Yahoo!

For this, the plan is to explore geometrical models of information retrieval [14] based on the theory of Quantum Mechanics and Hilbert Spaces, e.g. Tensor Space Models [7]. We believe, it is a promising alternative to leverage semantic metadata into the retrieval and summarization process.

Thereby, the suggested research is twofold:

1. to investigate the new geometry of IR and to exploit the possible ways of leveraging semantic metadata in the geometrical models of meaning and retrieval;
2. to explore the new ways of addressing users' information needs by means of summarization.

ACKNOWLEDGEMENTS

This work is supported by German Federal Ministry of Economics (BMWi) under the project THESEUS (number 01MQ07019).

REFERENCES

[1] Bonino D., Corno F. (2008) Self-Similarity Metric for Index Pruning in Conceptual Vector Space Models. In Proceedings of *19th International Conference on Database and Expert Systems Application (DEXA '08)*, pages: 225-229, Turin. IEEE Computer Society.

[2] de Rijke M., Meij E., Trieschnigg D. and Kraaij M. (2008) Parsimonious concept modeling. In *31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*, July.

[3] Gabrilovich E., Markovitch S. (2007) Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India.

[4] Gonzalo J., Verdejo F., Chugur I, Cigarran J. (1998) Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the COLING/ACL 98 Workshop on Usage of WordNet for NLP*, pages 3844, Montreal, Canada.

[5] Harris, Z. (1959). Linguistic transformations for information retrieval. In *Proceedings of the International Conference on Scientific Information*, Vol 2, Washington DC. National Academy of Sciences-National Research Council (NAS-NRC).

[6] Lin J. (2007) Is question answering better than information retrieval? Towards a task-based evaluation framework for question series. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 212219, Rochester, New York, April. Association for Computational Linguistics (ACL).

[7] Liu N., Zhang B., Yan j., Chen Z., Liu W., Bai F., and Chien L. (2005) Text representation: From Vector to Tensor. In Proceedings of ICDM, pages 725728. IEEE Computer Society.

[8] Luhn, H. P. (1958) The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2, 157-165.

[9] Mani I. (2001) *Automatic summarization.* John Benjamins Publishing Company.

[10] Moens M.-F. (2006) *Information Extraction: Algorithms and Prospects in a Retrieval Context.* The Information Retrieval Series.

[11] Nenkova A. (2008) Entity-driven rewrite for multidocument summarization. In *Proceedings of IJCNLP08.*

[12] Nenkova A., Vanderwende L., McKeown K. (2006) A Compositional Context Sensitive Multidocument Summarizer: Exploring the Factors That Influence Summarization. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 573-580. Association for Computing Machinery (ACM).

[13] Rose D. (2008) The information-seeking funnel. In *Position papers for the NSF Information Seeking Support Systems Workshop*.

[14] van Rijsbergen, C. J. (2004) *The Geometry of Information Retrieval.* Cambridge University Press, August.

An Analogy between the Double Slit Experiment and Document Ranking

Guido Zuccon
Dept. of Computing Science
University of Glasgow (Scotland)

Abstract

In the last years several works have investigated a formal model for Information Retrieval (IR) based on the mathematical formalism underlying quantum theory [1, 2, 3, 4, 5]. These works have mainly exploited geometric and logical–algebraic features of the quantum formalism, for example entanglement, superposition of states, collapse into basis states, lattice relationships. In this poster I present an analogy between a typical IR scenario and the double slit experiment. This experiment exhibits the presence of interference phenomena between events in a quantum system, causing the Kolmogorovian law of total probability to fail. The analogy allows to put forward the routes for the application of quantum probability theory in IR. However, several questions need still to be addressed; they will be the subject of my PhD research.

Keywords: document ranking, interdependent document relevance, quantum probability theory, interference, formal models

1. INTRODUCTION

Recently there has been an increasing interest on formal IR models inspired by the mathematical formalism of quantum theory. Some classical IR models, such as the vector space model, the probabilistic model, the logical model [3], and the logical imaging technique [5] have been expressed into the Hilbert space framework proper of quantum theory. In [4], the concepts of non-relevancy and negation of a term have been translated in orthogonality between subspaces of a Hilbert space. In [1], the role of context for word association is modeled by means of direct entanglement between words in high dimensional semantic spaces. Logic relationships between terms and the context surrounding them has been introduced in [2] by means of transformations on the text, the selective erasers, inspired by quantum measurements.

However, no previous work has been focusing on the nature of probabilities in IR. In particular, the use of quantum probability theory as feasible modeling tool for IR has never been explored. In the following, I introduce an analogy between the double slit experiment and a user examining a ranking of documents that have been retrieved by an IR system in response to his/her information need. The double slit experiment exhibits the arising of the interference phenomena; the presence of interference causes the violation of the Kolmogorovian law of total probability.

2. THE ANALOGY

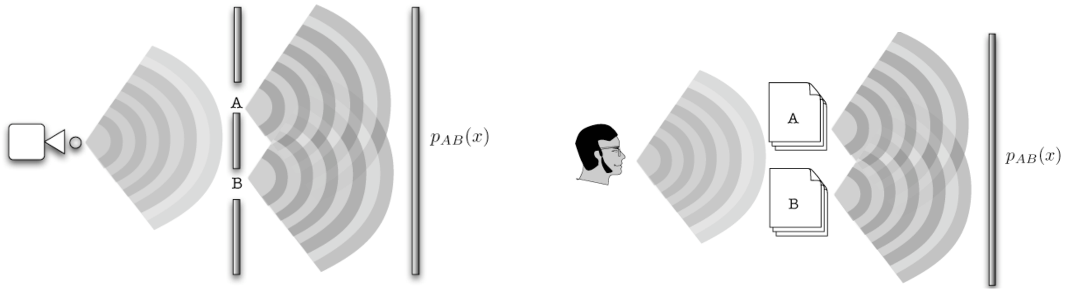
The double slit experiment consists of shooting a physical particle towards a screen with two slits, named A and B (Fig. 1(a)). Once the particle passes through one of the slits, it hits a detector panel, positioned behind the screen, in a particular location x with probability $p_{AB}(x)$. The (complex) probability amplitude associated to the events of passing through A (alternatively, B) when slit B (A) is closed and being detected at x is indicated by $\phi_A(x)$ ($\phi_B(x)$). Amplitude probabilities are linked to probabilities by the following equations: $p_A(x) = |\phi_A(x)|^2$, $p_B(x) = |\phi_B(x)|^2$. Intuitively, we would expect that the probability of the particle being detected at x when both slits are open is the sum of the probability of passing through A and being detected at x, $p_A(x)$, and the probability of passing through B and hit the detector panel in x, $p_B(x)$. However, experimentally it has been noted that $p_{AB}(x) \neq p_A(x) + p_B(x)$. Instead, the probability distribution obtained measuring $p_{AB}(x)$ across the detection panel presents an interference pattern akin to waves that would pass through both slits and hit the detector panel. In particular, the (complex) probability amplitude of a particle being measured at position x after passing through either slit A or B (indicated as $\phi_{AB}(x)$) is the sum of the probability *amplitude* associated to the event of opening just slit A plus the counterpart event of having open just slit B: $\phi_{AB}(x) = \phi_A(x) + \phi_B(x)$. In terms of probabilities,

$$p_{AB}(x) = |\phi_{AB}(x)|^2 = |\phi_A(x)|^2 + |\phi_B(x)|^2 + (\phi_A(x)^*\phi_B(x) + \phi_A(x)\phi_B(x)^*) = p_A(x) + p_B(x) + I_{AB}(x) \tag{1}$$

where term $I_{AB}(x)$ represents the quantum *interference* term, which is modulated by the phase difference between the amplitudes $\phi_A(x)$ and $\phi_B(x)$. In fact, by expanding $I_{AB}(x)$ and letting θ_{AB} being the phase difference between the probability amplitudes $\phi_A(x)$ and $\phi_B(x)$, we obtain

$$I_{AB}(x) = \phi_A(x)^*\phi_B(x) + \phi_A(x)\phi_B(x)^* = 2\,|\phi_A(x)|\,|\phi_B(x)|\cos\theta_{AB} \tag{2}$$

The analogy between the double slit experiment and the IR situation follows. The particle is associated with the user and his information need, while each slit represents a document (Fig. 1(b)). The event of passing through a slit is seen as the action of examining the ranking of documents, e.g. read the associated snippets or the documents themselves. Measuring at x stands for assessing the satisfaction of the user given the list of documents, or more concretely the decision of the user to stop his search (event x, the user is fully satisfied) or continue searching ($\bar{x}$, he is not completely satisfied by the documents). In these settings, being detected with probability $p_{AB}(x)$ at position x on the panel means choosing to stop the search with probability $p_{AB}(x)$ after being presented with documents A and B. Analogously to the double slit experiment, I propose that in the IR scenario the probability of the user being satisfied by the ranking of documents A and B is given by the sum of the probability of the single events (satisfied by document A, satisfied by document B) and the additional probability associated to the interference term. Thus, the satisfaction of the user does not depend just upon the relevance/satisfaction provided by each document independently (as it is commonly assumed in IR). Conversely, it is affected by the interference between the relevance/satisfaction of the entire document ranking. This suggests that a model of document ranking based on quantum probabilities might exploit interdependent document ranking.



(a) The double slit experiment

(b) The IR view of the double slit experiment.

FIGURE 1: Schematic representation of the analogy between the double slit experiment and the IR document ranking problem.

3. CONCLUSIONS

In this poster I have presented an analogy between the double slit experiment and a typical IR scenario. The analogy suggests that when calculating the probability of a document ranking being relevant to an user's information need the interference between the relevancy of the documents themselves should be accounted for. At this stage, several questions need to be investigated and they will be part of my PhD research. Among those, the most urgent are: What are the implications for IR? What do complex probabilities mean in IR? What does the interference term represent in IR? What is the behaviour of the interference term? How the interference term can be computed in IR?

4. ACKNOWLEDGEMENTS

This research has been sponsored in part by EPSRC project Renaissance (EP/F014384/1) and Royal Society International Joint Project JP080734.

REFERENCES

[1] P. D. Bruza, K. Kitto, D. L. Nelson, and C. L. McEvoy. Entangling words and meaning. In *Second Quantum Interaction Symposium*, 2008.

[2] A.F. Huertas-Rosero, L.A. Azzopardi, and C.J. van Rijsbergen. Eraser lattices and semantic contents: An exploration of the semantic contents in order relations between erasers. In *LNAI Series: LNCS, volume 5494*, pages 266–275. Springer Verlag, 2009.

[3] C. J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge Univ. Press, NY, USA, 2004.

[4] D. Widdows. *Geometry and Meaning*. CLSI Lec. Not. CSLI, 2004.

[5] G. Zuccon, L. A. Azzopardi, and C. J. van Rijsbergen. A formalization of logical imaging for information retrieval using quantum theory. In *IEEE Proc. 5th Int. Workshop TIR*. IEEE, 2008.

Eraser Lattices for Documents and Sets of Documents

Alvaro Francisco Huertas-Rosero
Uinversity of Glasgow
alvaro@dcs.gla.ac.uk

Abstract

Automatic schemes for the analysis of Natural Language based on word co-occurrence counting have been very successful in capturing meaning, like automatically grouping words referring to similar concepts, or documents about similar topics. In this work, a more general framework is proposed to represent documents and measurements geometrically, in a way directly related with the representation of measurement in Quantum Theory.

Keywords: Natural Language Processing, Quantum Logic, Lexical Measurements, Mathematical Models of Text

1. INTRODUCTION

Co-Occurrence of terms in text has been successfully used for automatically extracting semantic information from text documents (see [3], [4]). In this work, a different approach is proposed, based in transformations that act on documents in a way that is analogous to how projectors act on vectors. These transformations, called Selective Erasers, are defined in section 2. The underlying assumption behind this work is that suitably defined order relations between these measurement transformations are able to capture semantic contents of the text.

2. SELECTIVE ERASERS (SE) AND THEIR INCLUSION RELATIONS

A SE is defined as a transformations that find the occurrences of certain low-level feature in the document, preserve the surroundings, and erase the rest. This general definition is not very useful, because it does not specify what kind of low-level feature can be preserved *together with its surroundings*, and how these surroundings to be preserved can be defined. A more usable definition of a SE is given in [2] for the particular case of term occurrences (as low-level features) in text documents:

> A SE is a transformation $E(t, w)$ which erases every token that does not fall within any window of w positions around an occurrence of term t in a text document. These Erasers act as transformations on documents producing a modified document with some erased tokens, much as projectors act on vectors or other operators.

This concept was first introduced in [1], where some of their properties are shown, in particular, those they share with measurements as described in Quantum Theory. They can also be shown to include well known measurements such as occurrence and co-occurrences of terms and n-grams.

A very important characteristic is that some erasers will *include* others, which means that there will be pairs of Erasers such that what one preserves is included in what the other preserves. Each eraser will preserve small "windows" of text; when those corresponding to eraser A include within them those corresponding to eraser B, we can say that eraser A includes B *for the considered text*. The structure of these relations has been discussed in [2]. The formal condition for an inclusion relation between erasers (which will be denoted $E(t_1, w_1) \succ_D E(t_2, w_2)$ when it holds on document D) would be then:

$$E(t_1, w_1) \succ_D E(t_2, w_2) \iff E(t_2, w_2)E(t_1, w_1)D = E(t_2, w_2)D \tag{1}$$

3. FROM ERASERS TO PROJECTORS

Equation (1) defines a relation that is analogous to the inclusion relation between projectors on subspaces of a vector space. Changing SEs by projectors, and documents by vectors, the relations stand in the same way. The problem of representing Erasers and documents can be addressed through the following ansatz: **For a certain term t, the family of Erasers centred on it $E(t,w)$ would be accurately represented by a set of commuting projectors with rank $f(w)$, where f is a monotonic function**, This way relation $E(t,w) \succ E(t, w+\delta)$ are guaranteed for any integer, positive δ. The correspondence would be:

$$E(t,w) \equiv \Pi_{t,w} = \sum_{i=1}^{f(w)} |\psi_i\rangle\langle\psi_i| \text{ where for any two vectors } |\psi_i\rangle, |\psi_j\rangle \quad \langle\psi_i||\psi_i\rangle = \delta_{(i,j)} \tag{2}$$

Two projectors of the same rank corresponding to different central terms can be converted to each other by a unitary transformation, just like a term-swapping would convert the corresponding SEs:

$$E(A,w) \equiv \Pi_{A,w} = U_{(A \rightleftharpoons B)} \Pi_{B,w} (U_{(A \rightleftharpoons B)})^{\dagger} = \mathbb{T}_{(A \rightleftharpoons B)} E(B,2) \tag{3}$$

4. RANK OF PROJECTORS

A topic can be thought of as the set of documents about it, and can therefore be represented by inclusion relations. Suppose that for a document D_1 it is the case that $E(A,w_1) \succ_{D_1} E(B,w_2)$, and for document D_2, dealing with the same topic, it holds that $E(A,w_3) \succ_{D_2} E(B,w_4)$. The relation that holds *for both documents* would therefore be descriptive of the topic:

$$\begin{aligned}(E(A,w_1) \succ_{D_1} E(B,w_2)) \ \wedge\ (E(A,w_3) \succ_{D_2} E(B,w_4)) \\ \Rightarrow E(A, \max(w_1,w_3)) \succ_{\{D_1,D_2\}} E(B, \min(w_2,w_4))\end{aligned} \tag{4}$$

The increase in the width difference necessary to produce inclusion relations is crucial to determine the geometric representation of the Erasers. In the example of (4), the difference in width increases from $(w_1 - w_2)$ or $(w_3 - w_4)$ to $(\max(w_1,w_3) - \min(w_2,w_4)))$. Empirical evaluations shown in the figure suggest that this width can increase linearly with the number of documents considered.

To draw a vector analogy, we can consider the width factor of a SE can be considered as analogous to the rank of a projector. The join of two projectors will always include both of them, so join projectors can be related in this analogy to an including SE.

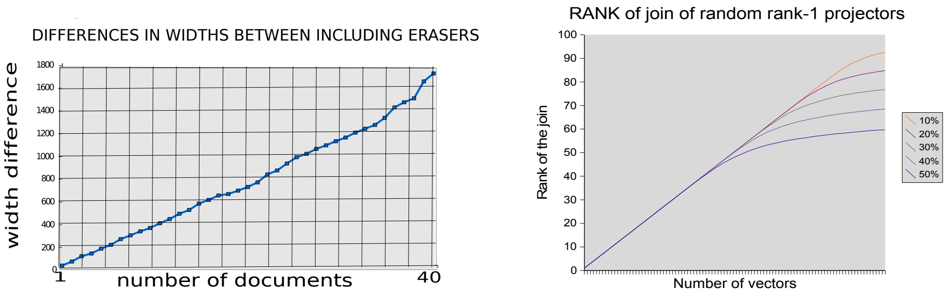


FIGURE 1: Measurement of widths required to produce inclusion relations, on approximately 2000 documents from TREC-1 that were assessed as relevant to 50 different topics. The linear increase of width suggest not to establish direct proportionality between width and rank of the corresponding projector. In the figure on the right, the average rank of joins made with random rank-1 projectors are shown. The different curves represent different threshold criteria to consider a vector as lying in a subspace (threshold ϵ for inner product). The less tight the threshold, the more the line gets closer to the straight line

Let us set a finite threshold for overlap $1-\epsilon$ to consider a unitary vector as lying in a subspace. Projector Π can be considered as the join of R disjoint 1-dimension subspaces, where R is its trace. A random rank-1 projector will only increase the rank if it is not included in any of these, and since these non-inclusions are independent events, the probability of increasing rank is the product of $R = Tr(\Pi)$ identical terms

$$P(\ Tr(\Pi \cup |\psi\rangle\langle\psi|) = Tr(\Pi) + Tr(|\psi\rangle\langle\psi|)\ ) = (1-\epsilon)^{Tr(\Pi)} \tag{5}$$

The curve showing the expected increase of rank with random vector in a space of dimension 100 is shown in figure 1 suggests that the dimension required to represent erasers as projectors is behind 40. The point where the curve starts showing a negative curvature, like that of the rank curves for projectors, will probably be only approached with bigger collections or more frequent terms. A closer study of this kind of curves could suggest which is the number of dimensions required to represent sets of Selective Erasers as projectors on subspaces of a Hilbert space.

5. ACKNOWLEDGEMENTS

This research has been sponsored by the Department of Computing Sciences of the University of Glasgow, as well as by EPSRC project Renaissance (EP/F014384/1) and Royal Society International Joint Project JP080734. The present work has also benefited from key ideas and guidance from prof. C. J. van Rijsbergen and Dr. Leif Azzopardi, from the Department of Computing Science of the University of Glasgow.

REFERENCES

[1] A.F. Huertas-Rosero, Leif Azzopardi, and C.J. van Rijsbergen. Characterising through erasing: A theoretical framework for representing documents inspired by quantum theory. In C. J. van Rijsbergen P. D. Bruza, W. Lawless, editor, *Proc. 2nd AAAI Quantum Interaction Symposium*, pages 160–163, Oxford, U. K., 2008. College Publications.

[2] A.F. Huertas-Rosero, Leif Azzopardi, and C.J. van Rijsbergen. Eraser lattices and semantic contents: An exploration of semantic contents in order relations between erasers. In C. J. van Rijsbergen P. D. Bruza, W. Lawless, editor, *Proceedings of the III Quantum Interaction Symposium QI2009*, volume 5494 of *Lecture Notes in Artificial Inteligence*, pages 266–275. Springer Verlag, 2009.

[3] Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, 1998. http://lsa.colorado.edu/papers/dp1.LSAintro.pdf.

[4] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instruments and Computers*, 28(2):203–208, 1996.