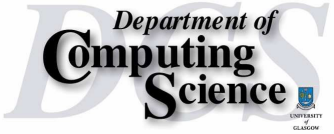
First International Workshop on Adaptive Information Retrieval (AIR)

University of Glasgow
Scotland, UK

14th October 2006

Abstract Booklet

Sponsors



First International Workshop on Adaptive Information Retrieval (AIR)

University of Glasgow
Scotland, UK

14th October 2006

Abstract Booklet

Editors: Hideo Joho, Jana Urban, Robert Villa, Joemon M. Jose, C.J. van Rijsbergen

Booklet Editor: Jana Urban

Sponsors

Multimedia Knowledge Management Network
Department of Computing Science
Adapt project (EPSRC Grant ref: EP/C004108/1)

Table of Contents

Invited Talk

Getting Personal: Personalization of Support for Interaction with Information

Nicholas J. Belkin
Department of Library and Information Science
Rutgers University
New Brunswick, NJ USA
nick@belkin.rutgers.edu

Abstract. One important aspect of adaptive information retrieval is *personalization* of the interaction with information to an individual's (or perhaps group's) context, situation, characteristics, and other factors. In this talk, I identify the goals of such personalization, discuss previous and current research in personalization, propose a classification of factors according to which personalization might be accomplished, and speculate on future research in personalization of interaction with information. I also discuss possible methods for large-scale, community-wide evaluation and comparison of personalization techniques.

Provocative Position Paper Session I

Simulating Searcher's Feedback, Quality and Effort in Interactive IR

Kalvero Järvelin

University of Tampere
Finland
kalervo.jarvelin@uta.fi

Abstract. Relevance feedback (RFB) is an important aspect of IR system adaptation to user needs. Experiments on the effectiveness of RFB with real users are time-consuming and expensive. This makes simulation for rapid testing desirable. We define a user model, which helps to quantify some interaction decisions involved in simulated RFB. First, we use the relevance threshold to model the user's acceptance of documents as relevant to his/her needs. Second, the browsing effort refers to the patience of the user to browse through the initial list of retrieved documents in order to give feedback. Third, the feedback effort refers to the effort and willingness of the user to provide RFB. We use the model to construct several simulated RFB scenarios in a laboratory setting. Using TREC data providing graded relevance assessments, we study the effect of the quality and quantity of the feedback documents on the effectiveness of the RFB and compare this to the pseudo RFB. Our results indicate that one can compensate large amounts of relevant but low quality feedback by small amounts of highly relevant feedback. They also suggest that IR system adaptation should be studied with graded relevance assessments: evaluation by liberal (TREC-like) relevance may hide important aspects of adaptation.

Issues in Research on Adaptive Search Systems

Joemon M. Jose

University of Glasgow, Department of Computing Science, 17 Lilybank Gardens, G12 8RZ
Glasgow, UK
jj@dcs.gla.ac.uk

Abstract. Adaptation of information retrieval systems is an important and popular research topic. I will discuss current approaches to the development of adaptive search systems in textual and multimedia retrieval domains. Such approaches vary in many respects – on the use of interface tactics and the retrieval models employed. Subsequently, I will elaborate current research methodologies employed for the evaluation of such systems and their limitations. Stumbling blocks in the development of such systems will be outlined.

Invited Talk

A Model of IR Testing and Evaluation: From Laboratory towards User-Involved

Noriko Kando

National Institute of Informatics (NII)
Tokyo, Japan
kando@nii.ac.jp

Abstract. Adaptive information retrieval probably has two sub-classes; collaborative adaptation by groups of users, and adaptation by single users within interaction or exploration. Either case, IR testing and evaluation methodologies and metric which have been widely used in the research and practice of the IR need to accommodate to the new environment. In this talk, for the first, I briefly introduce the activities of NTCIR, and then as an extension of these, I propose a model, or framework, of IR testing which covering from laboratory-type testings to user-involved tests in interactive setting and discuss about feasible strategies towards evaluation of adaptive information retrieval systems by step-by-step wise extension to the features related to adaptive.

Provocative Position Paper Session II

Test collections for all

Mark Sanderson

University of Sheffield
Sheffield, UK
m.sanderson@sheffield.ac.uk

Abstract. Researchers working in the IR field have placed a lot of reliance on building test collections that can be used widely by many researchers. Many collections have been used for years even decades. In the age of contextual IR, this talk will advocate an alternative far less tried approach, that of building many context specific collections, that don't require a great deal of effort to build but may not be all that re-usable. However, I shall argue that this is a better approach to take.

Employing User Relevance Assessments for Measuring Retrieval Effectiveness

David J. Harper[1]

[1] The Robert Gordon University
School of Computing, Aberdeen, Scotland, UK
d.harper@rgu.ac.uk

1 Background

In a recent user study [3], the TREC-8 Interactive Track collection was used [2]. This collection consists of a corpus of 210,158 articles from the Financial Times of London 1991-1994, a set of aspectual search topics, and a set of relevance judgments. The aspectual search task was first used in the TREC-5 Interactive Track [1]. This task was designed to mimic situations where users are not interested in finding all relevant documents on a particular topic, but instead are interested in finding documents that discuss different aspects or instances of a topic. Aspectual search topics used in the TREC-8 Track were created from six TREC Ad-hoc Track topics by adding a description field called 'instances' and removing the 'narrative', which provides guidance to users on judging relevance. One of the TREC-8 aspectual recall tasks entitled 'tropical storms' is displayed in Figure 1. A number of issues emerged while using this collection.

Number: 408i; **Title**: tropical storms

Description: What tropical storms (hurricanes and typhoons) have caused property damage and/or loss of life?

Instances: In the time allotted, please find as many DIFFERENT storms of the sort described above as you can. Please save at least one document for EACH such DIFFERENT storm. If one document discusses several such storms, then you need not save other documents that repeat those, since your goal is to identify as many DIFFERENT storms of the sort described above as possible.

Fig.1. Example TREC-8 Interactive Track topic

Recall Ceiling The original relevance assessments for TREC8 were based on pooled results from 7 participating groups [2], with a resulting document pool of just 1189 documents The relevance assessment task required that the TREC/NIST assessors both assess the relevance of each document against the topic, and identify relevant instances/aspects for each relevant document. Given the complexity of the assessment task, a relatively small pool of documents was assessed. When this collection is used in interactive retrieval experiments, it is likely that some retrieved documents may have no relevance assessments. In the study referred to above [3], we

observed a comparatively large number of such documents, namely 119 new documents, as shown in the final row of Table 1. Given that novel user interfaces are designed to enable alternative, or more thorough, explorations of the search space, this is likely to be the case with many such experiments. The usual approach adopted in (*ad hoc*) experiments using the TREC collections, is to assume that unassessed documents are not relevant. But, given that users have saved, and judged documents as relevant, this may result in misleading effectiveness measurements for interactive studies. That is, it is conceivable that some saved documents are relevant to the topic.

Judging Topic Relevance Subsequently, the authors of the above study used the original topic descriptions from TREC-8 to assess the unjudged 119 documents. They evaluated the documents independently, merged the evaluations, engaged in discussion of documents for which the assessments disagreed and arrived at a final judgment. Table 1 shows the distribution of unassessed documents across topic, and the total number of documents that we found relevant and not relevant for each topic. The levels of agreement in their original judgments for the four topics used in that study were: 428i (91%), 438i (83%), 431i (100%), and 408i (48%). We were surprised to see the low levels of agreement for some topics. It transpired that there was considerable scope for alternative interpretations of the topic, and we will focus here on just the most problematic topic, 408i (see Fig. 1). For 408i, the two judges disagreed on a variety of points, including what was meant by "damage", by "property", and what constituted a "different" storm, i.e. a storm that could be identified as a particular, and therefore different storm. In part, this difficulty arose, because of the way the TREC8 topics were derived. They were based on TREC *ad hoc* topics, but the 'narrative' field describing characteristics of relevant and non-relevant documents was removed. Similar observations concerning interpretation hold for the other topics, even though they proved less problematic in assessing relevance.

Table 1. Relevance assessments of documents saved in study [1], for which there were no TREC8 assessments. R: relevant; NR: not relevant.

	Topic				Total
	408i	**428i**	**431i**	**438i**	
R	12	3	0	9	24
NR	19	20	19	37	95
Total	31	23	19	46	119

TREC Assessor Judgement Given the differences between two judges described above, it is reasonable to ask, on what basis did the TREC/NIST assessors make their relevance assessments. It seems clear that the NIST assessors would face the same problem as any user, namely how to interpret the topic. Potentially, this was further complicated by the fact that, some assessors were the *originator* of the ad hoc topic [2], for which they had provided original 'narrative' text, and this may have affected their judgments. Certainly, the NIST assessor would have settled on a particular interpretation. At this point, one might well ask: why is the NIST assessors' interpretation, in the "absence" of a narrative field, now accepted as the "gold

standard" when assessing retrieval effectiveness using this collection? Why are the relevance assessments made by users participating in user studies any less valid, or put another way, why is their topic interpretation any less valid[1].

In the rest of this paper, we will explore the idea of directly using end user assessments of relevance in determining retrieval effectiveness, and will discuss some of the implications of this approach.

2 Topic Interpretation and Effect on User Studies

Given the nature of the TREC8 interactive track topic/task, there is clearly scope for differing interpretations of topicality, and what is relevant. We conjecture there will be a core set of documents, on which most users will agree are relevant, and another set for which users disagree due to differences in interpretation. In this respect, NIST assessors will be no different, sharing some common parts of the generally applicable interpretation. The problem is if any given user differs in their interpretation from the NIST assessor, then the performance measured for this user when using the NIST assessments, could be substantially reduced, even though the interpretation may be a reasonable one. When we measure the Precision of any a particular search by the user, it may be that what we are measuring is (in large part) simply the agreement between the user and the assessor on the interpretation. Precision results from a range of user studies consistently show average values around 0.6-0.8 [2] [3], and indeed inter-judge agreement on relevance assessment tasks, has been shown to be comparatively low [4]. We conjecture that any differences between systems under study *may* be masked by the differences in effectiveness due to these differing interpretations. That is, there may be greater variation in the effectiveness measures, making it harder to demonstrate statistically significant differences between systems under test, where these exist. The exclusion of the narrative section from the topics, may have introduced a potentially confounding variable, namely topic interpretation. Our idea is to embrace these differing interpretations, and to directly use the (pooled) user assessments in measuring retrieval effectiveness.

3 Measuring Effectiveness with User Relevance Assessments

Suppose we conduct a user study, in which documents are examined, and saved (or not) depending on whether the document is (or is not) relevant for the topic[2]. For each topic, for each document saved by at least user for the topic, we will have the following data: number of users who saved the document, and number of users who

[1] Granted, some users certainly adopt questionable interpretations of a topic, and/or simply make mistakes. For example, some users saved documents describing "hurricane" force winds in the UK and Italy, as relevant to the topic 408i!

[2] To simplify the discussion, we will consider just the simple binary assessment of relevance, without considering the complication of determining aspects/instances.

displayed (say) the document, but did not save it. (We will ignore whether the user viewed a snippet in a result list). The higher the ratio between saved/not saved for a given document, the higher the probability that the document is relevant to the topic, or at least relevant in respect of the core interpretation. We could use such a ratio (or probabilistic equivalent) to determine a plausible set of relevant documents for a topic, based on the user judgments. We could formulate this in probabilistic terms as estimating the probability of *assessing* a document as relevant, give a particular document, topic and set of users. We will denote this *P(assessRel)*, but will not describe how to estimate this in the paper.

How then might we use the save/not saved ratio or probability, *P(assessRel)*, for measuring retrieval effectiveness? For convenience, we will refer to *P(assessRel)* in the following. We could rank the saved documents in decreasing order by this probability, apply a cutoff, and deem documents above the cutoff as relevant. In essence, we would be establishing a common interpretation, and core set of relevant documents. Alternatively, we might try and establish a three way partition of the documents: relevant, not relevant, and "open to interpretation". In Figure 2, we show a likely typical distribution of P(assessRel) for a set of documents saved for a topic by a group of users. We can identify three regions, corresponding to the three-way partition referred to above. The middle partition comprises those documents for which there is considerable disagreement about the assessment, i.e. those for which a comparable number of users judge a document relevant or non-relevant[3]. Then, we might choose to ignore, or at least reduce the influence of, the documents in the "open to interpretation" partition when computing effectiveness measures. A third way would be to use the *P(assessRel)* estimates directly in probabilistic variants of effectiveness measures, such as probabilistic Precision and Recall measures, where simple counts of relevant documents are replaced by sums over *P(assessRel)*. No further details will be provided due to space constraints.

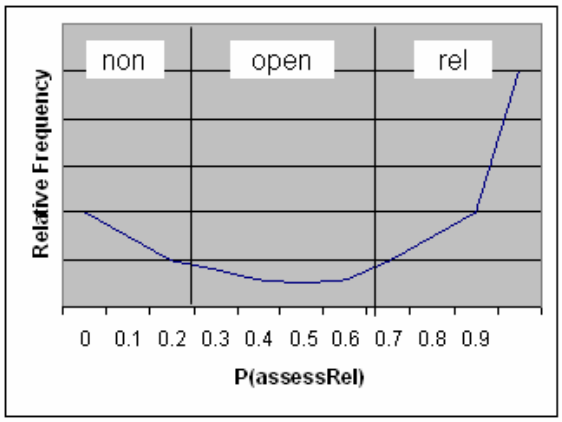


Figure 2: Representative Graph of Relative Frequency of Saved Documents against P(assessRel) for a Topic.

[3] Note, that for a "simple" topic having only a single interpretation, one might expect a large spike at the right end of the graph, and a somewhat smaller spike at the left, corresponding to documents for which a user has likely judged a document relevant in error.

4 Discussion and Conclusions

In this short paper, we have argued that for interactive information retrieval studies, where actual users are providing relevance assessments, we should consider using their pooled assessments when evaluating the effectiveness of the systems under study. Currently, it is usual to evaluate effectiveness using the NIST assessor judgments as the "ground truth", when in fact these judgments correspond to just one interpretation of the topic. We have outlined a way of using the pool of user assessments to estimate the probability that a (saved) document is assessed relevant, given the document, topic and set of users. Further, we have proposed three ways in which these estimates might be used in measuring retrieval effectiveness. This proposed methodology raises a number of interesting questions itself.

In computing *P(assessRel)*, what group of users should we use when establishing the pool of assessments? The obvious pool is that derived from the study itself, and this may be a good choice given that (a) we wish to compare the systems that generated the pool, and (b) we are generally unable to compare retrieval effectiveness across user studies due to, among other things, differences in the user populations. However, the pool from a single user study is comparatively small, and it is worth considering pooling assessments from a number of studies, in order to provide a more representative pool. This leads to some interesting questions about how to establish a stable test collection, so that experiments can be replicated. Further, the proposed methodology is based on saving (or not) documents, and inferring the relevance assessment from this behaviour. We would urge those who conduct such studies in the future to log both saved documents, and viewed documents, to provide information for computing *P(assessRel)*. This technique only provides information about the assessed relevance of a document, and does not provide information about relevant instance/aspects as per the TREC8 interactive framework. Instance assessment would need to be performed by a human judge. But, we could likely reduce the number of documents to be judged to the probably relevant ones.

The kind of graph shown in Figure 2 might prove a useful tool in evaluating and understanding studies of interactive information retrieval. Given such a graph, we may be able to measure or estimate the degree to which a topic admits of multiple interpretations, by seeing the extent to which the distribution is skewed towards either end, i.e. "spiky at the ends", or not. This could be used to explore the effect of including a narrative or not, and the effect of task on relevance assessment. It might be used for looking at the assessment behaviour of individual users in studies, to identify users whose assessments lie outside the usual range.

The paper has also pointed out that the recall ceiling of the TREC8 Interactive Track collection is comparatively low, and experimenters should be aware of this, and prepared to do something about retrieved documents, for which there are no assessments.

We have implicitly assumed that an experimenter will want to measure retrieval effectiveness when undertaking interactive studies. Clearly, there are a range of other types of interesting measurements possible, including measures of user satisfaction, measures of interaction, including time for task, and other newer measures such as cognitive load. It will depend on the research questions being explored, as to the most appropriate measures for a given study.

Finally, we would like to state that this paper is not intended as a criticism of the design of the TREC8 interactive track experiment, or indeed the general TREC approach of relevance assessment. The interactive track experiment was designed so that users would interact vigorously with systems under study, and the aspectual retrieval task is highly successful in this regard. This paper does point out some limitations of the TREC8 interactive test collection, and proposes some alternative ways of thinking about measuring retrieval effectiveness, where the users' assessments of relevance are employed.

Acknowledgements I would like to thank both Paul Over and Bill Hersh who shared their knowledge of the TREC8 track with me. Naturally, the opinions expressed here about TREC8 are my own. I would also like to thank Diane Kelly for the interesting discussions that ensued when we judged the relevance of unassessed documents for the study [1], even though I still think 'damage to property' should include 'damage to crops' for topic 408i.

References

1. Dumais, S. T., & Belkin, N. J. (2005). The TREC Interactive Tracks: Putting the user into search. In E. M. Voorhees & D. K. Harman (Eds.) *TREC: Experiment and Evaluation in Information Retrieval* (pp. 123-153), Cambridge, MA: MIT Press.
2. Hersh, W., & Over, P. (1999). TREC-8 interactive track report. In D. Harman and E. M. Voorhees (Eds.), *The Eighth Text Retrieval Conference (TREC-8)*, 57-64.
3. Harper, D J., & Kelly, D. (2006). Contextual Relevance Feedback. In *Proceedings of 1st Symposium on Information Interaction in Context*, Copenhagen, Denmark, October 2006 (to appear).
4. Vorhees, E. M. Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness. In *Proceedings of 21st ACM SIGIR Conference*, Melbourne, Australia, 315-323.

Invited Talk

Building Test Collections for Adaptive Information Retrieval: What to Abstract for What Cost?

Ellen Voorhees

NIST
ellen.voorhees@nist.gov

Abstract. Traditional Cranfield test collections represent an abstraction of a retrieval task that Sparck Jones calls the "core competency" of retrieval: a task that is necessary, but not sufficient, for user retrieval tasks. The abstraction facilitates research by controlling for (some) sources of variability, thus increasing the power of experiments that compare system effectiveness while reducing their cost. However, even within the highly-abstracted case of the Cranfield paradigm, meta-analysis demonstrates that the user/topic effect is greater than the system effect, so experiments must include relatively large number of topics to distinguish systems' effectiveness. The evidence further suggests that changing the abstraction slightly to include just a bit more characterization of the user will result in a dramatic loss of power or increase in cost of retrieval experiments. Defining a new, feasible abstraction for supporting adaptive IR research will require winnowing the list of all possible factors that can affect retrieval behavior to a minimum number of essential factors.

Posters

1. **Shi-Ming Huang, Chih-Fong Tsai, Chia-Ming Chuang and John Tait**, "Learning Users' Searching Behaviour in Image Retrieval"
2. **Peter Bailey and George Ferizis**, "Possible Approaches to Evaluating Adaptive Question Answering for Mobile Environments"
3. **Giridhar Kumaran and James Allan**, "Eliciting Information for Adaptive Retrieval"
4. **Udo Kruschwitz and Maria Fasli**, "Adaptive IR via Automatically Maintained Domain Knowledge"
5. **Hyowon Lee, Alan F. Smeaton, Noel E. O'Conner and Gareth J.F. Jones,** "Adaptive Visual Summary of LifeLog Photos for Personal Information Management"
6. **Frances C. Johnson**, "Optimal Results Presentation for Dynamic Search"
7. **Giorgio M. Di Nunzio and Nicola Ferro**, "Queries and Relevance Assessments: The Right Context for the Right Topic"
8. **Frank Hopfgartner, Robert Villa and Jana Urban**, "Adaptive Video Retrieval"
9. **Hideo Joho and Joemon M. Jose**, "Effectiveness of Additional Representations in the Search Result Presentation on the Web"
10. **Hideo Joho and Joemon M. Jose**, "Using Local Contexts to Slice and Dice the Search Results on the Web"
11. **Robert D. Birbeck, Hideo Joho and Joemon M. Jose**, "A Sentence-based Ostensive Browsing and Searching on the Web"
12. **Andres Masegosa, Hideo Joho and Joemon M. Jose**, "Identifying Features for Relevance Web Pages Prediction"
13. **Ioannis Psarras and Joemon M. Jose**, "A System for Adaptive Information Retrieval"
14. **Jana Urban**, "Combining Image Organisation and Retrieval to Overcome the Semantic Gap in CBIR"
15. **Sachi Arafat**, "A Formal Approach to Information Retrieval based on Quantum Mechanics"

Learning Users' Searching Behavior in Image Retrieval

Shi-Ming Huang[1], Chih-Fong Tsai[2], Chia-Ming Chuang[1], and John Tait[3]

[1] Department of Information Management, National Chung Cheng University, Taiwan
[2] Department of Accounting and Information Technology, National Chung Cheng University, Taiwan
[3] School of Computing and Technology, Sunderland University, UK
smhuang@mis.ccu.edu.tw; actcft@ccu.edu.tw; jason041167@gmail.com; john.tait@sunderland.ac.uk

Abstract. Relevance feedback is able to improve retrieval effectiveness in content-based image retrieval. However, it may be a tedious task for users to provide a number of positive and/or negative relevance indicators for the retrieved images. This paper presents an Enhanced Semantic-Based Mechanism (ESBM) which uses semantic categories (such as *pretty*, *peaceful*, etc.) based on color as the initial keyword-based queries, collaborative filtering to group users into several clusters based on users' searching behavior or historical data and 'implicit' feedback as the searchers' clicks to download some of the retrieved images. The proposed approach outperforms systems using relevance feedback and collaborative filtering separately.

Keywords: content-based image retrieval, relevance feedback, collaborative filtering

1 Introduction

The performance of Content-based Image Retrieval (CBIR) systems is unsatisfactory for many practical applications due mainly to the semantic gap between searchers' high-level conceptualization of their query and the low-level visual features of images (Jörgensen, 2003).

One way to solve the semantic gap problem is to use relevance feedback during image retrieval (Zhou and Huang, 2003). Following a number of relevance judgements provided by searchers, systems are able to synthesize queries which retrieve more relevant images compared to the retrieval results of the initial query. However, users need to provide numbers of judgements to improve system's performance. In addition, it is not known how many feedback iterations are required to reach searchers' ideal.

We propose an Enhanced Semantic-Based Mechanism (ESBM), which uses relevance feedback, collaborative filtering, and color-based semantics (i.e. semantic categories, such as 'pretty', 'cheerful', 'modern' based on colour signatures), to adapt the image indexing and retrieval process in order to overcome the semantic gap and improve retrieval effectiveness and user satisfaction (Chuang, 2005; Huang et al., 2006).

2 Experimental Results

We used mobile images at the Samsung Chinese website[1] for our experiments. The images at this website are classified into 15 categories and there are 592 images. Sixty general searchers who do not have image retrieval background were asked to participate in this system evaluation. First of all, thirty searchers were asked to query images and download some of the retrieved images if they wished. Then, the searchers' personal information was clustered and the ranking score of images was obtained based on their searching behavior (i.e. the queried keywords and downloaded images as the implicit feedback).

Figure 1 shows the results of retrieval accuracy of ESBM and two systems in which system 1 uses relevance feedback only and system 2 uses collaborative filtering only. On average, ESBM, system 1 and 2 produce the avg. accuracy of 85.86%, 65.59%, and 70.66% respectively. Moreover, ESBM significantly outperforms both system 1 and system 2 (p>0.99) and system 2 outperforms system 1 (p>0.99).

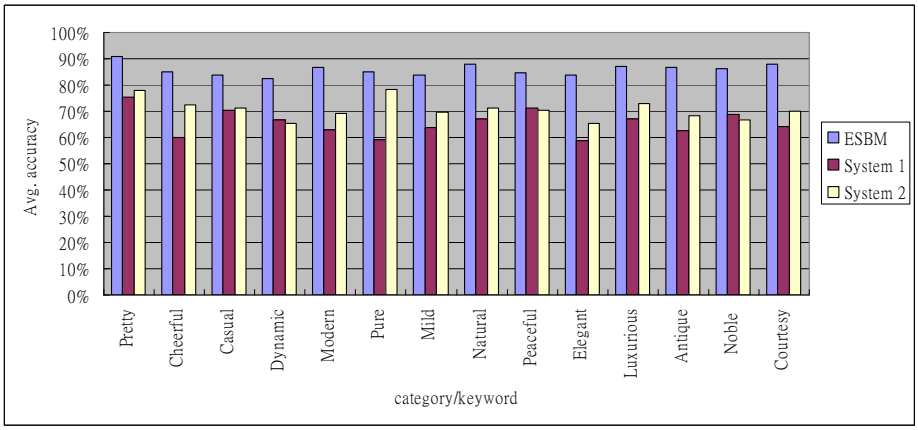


Fig. 1. Avg. accuracy of the three systems.

In summary, combining relevance feedback, collaborative filtering, and color-based semantics to adapt the search and retrieval process improves retrieval accuracy.

References

1. Chuang, C.-M.: An Enhancement Semantic-Based Mechanism for Image retrieval. Master Thesis, National Chung Cheng University, Taiwan (2005)
2. Huang, S.-M., Tsai, C.-F., Chuang, C.-M., and Tait, J. Learning Users' Searching Behavior in Image Retrieval. Pattern Recognition (under review)
3. Jörgensen, C. (ed.): Image Retrieval: Theory and Research. Scarecrow Press, Oxford (2003)
4. Zhou, X.S. and Huang, T.S.: Relevance feedback in image retrieval: a comprehensive review. Multimedia Systems 8(6) (2003) 536–544

[1] http://hk.samsungmobile.com/ct/play/color_images/prm_mw_list.jsp?p_pfid_1=&p_cate=125&inquiry=&order=datemodified

Possible Approaches to Evaluating Adaptive Question Answering Systems for Mobile Environments

Peter Bailey and George Ferizis

CSIRO ICT Centre, GPO Box 664, Canberra ACT 2601 AUSTRALIA
{Peter.Bailey, George.Ferizis}@csiro.au

Introduction

The CSIRO ICT Centre has recently constructed a question answering (QA) system – My Instant Expert™ – designed for mobile phones. The client-server system supports asking open domain natural language questions and attempts to find answers from the (English) Wikipedia. Due to the small display of mobile phone devices, the space available for both question entry and answer display is limited (e.g. 240x320 pixels).

Adaptive IR and delivery techniques appeal as methods to maximize the use of this limited display and minimise the use of the costly and low network bandwidth. QA systems typically are constructed from a mixture of information retrieval (IR) and computational linguistics technologies. Adaptive approaches to IR are posited as being more likely to improve overall user satisfaction with performance. It is unclear that the existing format of QA test collections will work effectively for evaluation.

Experience

There is substantial related work in the area of building QA test collections, for example [5]. In the iterative prototyping development of the system, we faced the problem of not having a reliable baseline to benchmark our work against. Our compromise was to use a large selection of questions from TREC QA track topics, then manually verify that identical answers to these questions existed within the English Wikipedia corpus. TREC QA track-style answer patterns were used to identify whether retrieved answers contained matches.

The system uses a fairly standard approach to pipelining a sequence of IR and computational linguistic components. Performance at each component's output stage was measured using standard metrics. Representativeness of these TREC questions was an issue, especially with respect to having few questions with numeric/scale answer types. This approach provided us with some measure of server-side performance of the system, but specific to the exact TREC question set.

A significant issue with deploying a QA system on a mobile phone is the user experience of interaction, including answer presentation and answer-in-context display. In other words, the whole client-side of the equation is important to consider.

Possible approaches to evaluation

Our experiences and the additional challenges of adaptivity lead us to support the directions recently articulated by Sparck Jones's [3]. Her analysis framework which captures input, purpose, and output factors is more appropriate for adaptive systems.

Specifically, we believe it is vital to consider, model and assess output factors such as format and brevity when the interaction device is a mobile phone. Similarly, input factors such as the form of the source and subject type (e.g. news articles vs Wikipedia articles) play a part in understanding how users will interact with and trust the information. Most importantly, purpose factors such as audience and use (e.g. answering trivia questions is different from answering current stock prices) are essential for evaluating the quality of a QA system.

A comparative system evaluation approach. When the purpose of evaluation is to improve system performance and/or user satisfaction, not to compare it to past performance of other systems, then test collections need not be reusable. The approach of Thomas and Hawking involving side-by-side comparative judging in context of result displays is appropriate here [4].

To provide support for repeated evaluations, elements of standard test collections can be valuable. These include a set of queries, preferably a larger set of queries than usual, and they should be real user queries. To support this, we intend to provide a substantial query log from the My Instant Expert™ system in the coming months. Similarly a fixed corpus such as the INEX Wikipedia corpus [1] is preferred.

The classic test collection approach, updated. If there are many groups working on the same aspect of adaptive QA, then the creation of a reusable QA test collection becomes more valuable. The approach of [2] to address the creation of reusable answer judgements for specific sub-problems (e.g. factoid questions) could be adopted, with appropriate sampling and search-mediated judging. Making the test collection reusable will be far more complex, as it will entail modeling the additional factors appropriately. For example, if the query is “how many players are there in a football team?”, and if input factor locality is “UK” then the answer is 11 (soccer); if input factor locality is “New Zealand”’ then the answer is 15 (rugby union).

References

1. Denoyer, L. and Gallinari, P. 2006. The Wikipedia XML Corpus. *SIGIR Forum*
2. Lin, J. and Katz, B. 2006. Building a reusable test collection for question answering. *J. Am. Soc. Inf. Sci. Technol. 57*, 7, 851-861
3. Sparck Jones, K. 2006. What's the value of TREC: is there a gap to jump or a chasm to bridge? *SIGIR Forum 40*, 1, 10-20.
4. Thomas, P. and Hawking, D. 2006. Evaluation by comparing result sets in context. In *ACM Fifteenth Conference on Information and Knowledge Management* (November 2006).
5. Voorhees, E. M. and Tice, D. M. 2000. Building a question answering test collection. In SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New York, NY, USA, 2000), pp. 200-207.

Eliciting Information for Adaptive Retrieval

Giridhar Kumaran and James Allan

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
140 Governors Drive, Amherst MA 01003 USA

1 Introduction

The task of developing interaction strategies [1] involves determining what additional information will be useful in the context of the query, and the method to obtain this information. In the quest to obtain as much data from the user as possible it is important to keep usability in mind. The most complex interaction mechanisms, however effective, can discourage a user due to high cognitive load. This motivates us to focus on developing a suite of very effective interaction strategies that do not demand much effort, cognitive and physical, from the user. User responses are aimed to be simple too - usually yes/no decisions or selecting from a very small set of options. While this explorative study did not involve an actual user study, each of the techniques described have the potential to be more effective in an interactive setting.

2 Simple Techniques

We designed a few interaction strategies to handle a subset of failures described in a study of why search engines fail [3].

1. *Spelling mismatch due to typographical errors and cultural differences.* To address this problem, we used string edit distance as a simple type of spelling correction, and treated the variants found as synonyms. The user could be asked to verify if the identified variant was truly one.
   Is oestrogen a reasonable variant spelling of estrogen?
2. *Recognizing phrases in the query using punctuation.*Apostrophes, hyphens and double quotes which are usually discarded while indexing indicate the possibility that the associated terms form a phrase. For example, in response to the query *Find documents that discuss issues associated with* **so-called "orphan drugs"**, a user could be asked
   Is it correct that you see so called as a phrase related to the query?
   Is it correct that you see orphan drugs as a phrase related to the query?
3. *Identifying patterns in top-ranked documents* Similar patterns of terms, either as phrase or within certain term windows, occur frequently in similar documents. Questions posed to the user could be of the form
   Would you expect to see leaning and pisa nearby, with terms such as tower and of between them?

3 Interesting Directions and Challenges

Experiments with the three questions described in the previous section with **simulated** interaction[1] on the TREC 2004 and 2005 Robust track data sets have validated their utility[2] . We are currently looking at several additional interaction strategies, mostly motivated by the availability of data annotations from the Automatic Content Extraction program.

1. *Entity context.* It is useful to have a mechanism to further clarify the context a term or entity is used in. For example, users can define context by reporting if the term 'Bonaire' should be part of an address,(*Bonaire, Netherlands Antilles*) or an organization (*Bonaire Democratic Party*).
2. *Person named entities.* The user can be asked to choose the entities related to the query found in the top-ranked results from an initial run. A very short biography from a source like Wikipedia can help the user make the decision.
3. *Top-ranked sentences.* The negative feedback obtained by asking the user to mark non-relevant sentences from the top-ranked ones could be used to clear the results list or reformulate the query.
4. *Targeting named entities.* Specifying the type of named entities the user is interested in can help disambiguate and focus a query.
5. *Query expansion/relaxation.* Providing users with pictorial feedback in the form of an online pie chart showing the percentage of the corpus affected by addition or removal of query terms could potentially guide the user in determining the best set of terms to use in a query.

Each of the above interaction strategies are *light-weight*, but in unison could defeat our goal of minimal interaction. Determining a set of appropriate strategies on a per-query basis is a challenge, with implicit feedback playing a major role. Adapting for different environments - the web, TREC-style querying or templated querying - is a challenge too. In addition to using IR metrics like precision and recall for evaluating result quality, we plan to develop or adapt measures from other areas to measure aspects like cognitive load and usability.

Acknowledgments This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

References

[1] Kumaran, G., Allan, J.: Simple Questions to Improve PseudoRelevance Feedback Results. ACM SIGIR 2006 Proceedings (2006) 661–662

[2] Harman D., Buckley C.: The NRRC reliable information access (RIA) workshop ACM SIGIR 2004 Proceedings (2004) 528–529

[1] Although our experiments have sidestepped the actual questions, we envision each of the techniques being used interactively

[2] The techniques apply to a small subset of the queries, which were improved on precision

Position Paper: Adaptive IR via Automatically Maintained Domain Knowledge

Udo Kruschwitz and Maria Fasli

Department of Computer Science, University of Essex, Colchester, CO4 3SQ, U.K.

1 Introduction

A massive number of electronic document collections exist within companies, universities and other institutions. However, locating relevant information within such collections can be difficult. Nevertheless, all of these collections do contain a huge amount of valuable knowledge that is encoded implicitly and can therefore not be applied directly. A challenging issue is to first identify and extract such knowledge automatically and then make it usable by incorporating it in a search system that assists users who want to search or explore the document collections. A search engine that does not simply return the results but instead offers the user suggestions to widen or narrow down the search has the potential of being a much more useful tool, e.g. [1]. How can such knowledge help a user in the search process? A student who searches a university Web site for "exam results" for example may be presented with a list of module names or numbers to choose from to narrow down the search. These query modification options would be constructed based on what is encoded in the knowledge derived from the documents. Automatically constructed knowledge can however never be as good as manually created structures. Therefore an even bigger challenge is to improve and maintain this knowledge - again automatically.

2 Research Outline

Different techniques exist to extract tree-structured domain knowledge for document collections, e.g. [2–4]. Any such domain model can be incorporated in a standard search engine to suggest query modification terms to the user in an interactive search process. But to our knowledge there is very little work (in fact almost no work) on updating/adjusting/adapting/evolving such a domain model based on either explicit or implicit user feedback. As an automatically extracted domain model will inherently be incomplete and contain a lot of "noise", adjusting the domain model is essential if the recommendations provided by the system are to be improved. Adapting the domain model is required in particular in situations where the pool of documents is not static, but dynamic. Continuously recreating the domain model seems inappropriate as there is always the question of how often this should be done, and moreover the newly created domain model will again be incomplete and contain a lot of noise. Instead we require a more flexible method that will enable us to filter out this noise from the domain

model. The hypothesis is that the users' search behaviour can be used as input into this process of adjusting the domain model so that it becomes more accurate. We are focusing on a specific aspect of this search behaviour, namely the selection of query modification terms which provides us with *implicit feedback* from the users and should be sufficient to come up with a model to automatically adjust the domain knowledge without having to rely on other forms of explicit or implicit user feedback [5]. However, our use of implicit relevance feedback is different from previous approaches in that we do not utilise it in a particular search task but instead we collect the feedback of the entire pool of users of the system in order to automatically adjust the domain model. In essence, we observe the behaviour of the user population and thus improve the domain model in a collaborative way. We also want to stress that our aim is *not* to build up individual user profiles which is a whole research field on its own, e.g. [6].

In order to devise a solid methodology for evolving automatically derived domain knowledge we require real user data. In this context we are not interested in general Web search. Therefore, we need user data for different collections. We have made a start by running a prototype of our own search system that combines a standard search engine (in our case *Nutch*) with automatically extracted domain knowledge [4]. The system has been running since late May on the University of Essex intranet. This allows us to collect a corpus of user queries (about 100 queries per (week)day; more than 6000 in total so far), interactions with the search system and, most importantly, click-through data such as information about what query modifications the users choose to select or construct and which suggestions the users tend to ignore. The log files we keep collecting are an extremely valuable resource because they are a reflection of real user interests (different to TREC like scenarios which are always a bit artificial). Nevertheless, it is also more difficult to interpret what the user was actually after. The data collected so far are a justification for a system that guides a user in the search process: more than 10% of user queries are query modification steps, i.e. the user either replaces the initial query or adds terms to the query to make it more specific. About three quarter of these modifications are terms suggested by the system (the others are additional query terms provided by the user).

References

1. Kruschwitz, U., Al-Bakour, H.: Users Want More Sophisticated Search Assistants - Results of a Task-Based Evaluation. JASIST **56**(13) (2005) 1377–1393
2. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In: Proceedings of SIGIR. (1999) 206–213
3. Anick, P.G., Tipirneni, S.: The paraphrase search assistant: terminological feedback for iterative information seeking. In: Proceedings of SIGIR. (1999) 153–159
4. Kruschwitz, U.: Intelligent Document Retrieval: Exploiting Markup Structure. Volume 17 of The Information Retrieval Series. Springer (2005)
5. Ruthven, I., Lalmas, M.: A survey on the use of relevance feedback for information access systems. Knowledge Engineering Review **18**(2) (2003) 95–145
6. Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: Proceedings of SIGIR. (2005) 449–456

Adaptive Visual Summary of LifeLog Photos for Personal Information Management

Hyowon Lee, Alan F. Smeaton, Noel E. O'Connor and Gareth J.F. Jones

Adaptive Information Cluster & Centre for Digital Video Processing
Dublin City University, Glasnevin, Dublin 9, Ireland
hlee@computing.dcu.ie

As the cost of taking a photo drops, due to the use of digital cameras, and the incentive to take photos increases, due to the ubiquity of camera phones and easy ways to share them with a large number of viewers, the average number of personal photos produced has increased dramatically. The need for supporting *access* to a large number of photos by efficient browsing and searching has become more crucial than ever, and the use of context information such as GPS location and time of capture are currently being researched to reduce the user's annotation burden and to aid retrieval.

Enter *passive capture* - the user attaches a camera the size of a button on her chest. The camera automatically and regularly takes photos whenever an interesting event happens throughout the day while the user goes about her daily activity or holiday trip. The *SenseCam*, developed by Microsoft Research, is a small digital camera that a user wears around her neck. It contains a number of sensors including infra-red and motion sensors to automatically trigger photo capture in such a way that the photos taken are not blurred. On a typical day, the SenseCam will take 3,000 - 4,000 photos throughout the day capturing meaningful and significant images of the wearer's activity, in effect chronicling most of the day's events.

At the end of the day, the wearer can download all the photos from the SenseCam to their computer as a detailed visual record of the day. The fact that everything is captured for reviewing or searching is comforting on the one hand, yet going through over 3,000 photos for each day can take a long time and much effort, and when multiple days are captured it becomes prohibitive to try to extract anything of use from this number of photos.

We use various content-based image analysis techniques on SenseCam photo collections spanning multiple days to automatically detect visual events. For each event we detect a landmark photo as a kind of "keyframe". We then automatically compose an interactive browser that summarises, emphasises and can replay thousands of SenseCam photos on a *single page* in an efficient and comfortable way so as to not overload the viewer (see Figure 1). Significant events are detected among each day's photos and their uniqueness or importance scores are calculated by examining how frequently and for how long similar events have occurred during the previous 1-week period. For example, the wearer working in front of a computer for 2 hours in the morning would appear almost every day, and thus such a visual event scores as less important; whereas a 15-minute unexpected meeting with a colleague on a corridor which happened only once in

the whole week, is given a higher importance score. The first day's desk work in a different university lab for a research visit would be determined as an important event as this is unique among the visual events of the preceding week, but as the days pass the desk work at the same lab will bear less and less importance as it becomes a common activity. Thus, the system adaptively re-ranks the importance of each event as the day's photos come into the database using a 7-day window. The current day's photos are presented as a comic book style interface with different size photos according to their ranked importance.

At the poster session, the overall information flow and processing of images will be presented with the SenseCam device and a few sample collections of its photos. The prototype of the interactive browser under development will be presented which dynamically summarises a large number of photos in a highly inviting, simple, and enjoyable way.

Figure 1. Interactive browser for reviewing a day's SenseCam images

Acknowledgments

This work is supported by Science Foundation Ireland under grant number 03/IN.3/I361, by Microsoft Research and by the EC under contract FP6-027026 (K-Space). We are grateful to the aceMedia Project for use of the aceToolbox.

Optimal results presentation for dynamic search (a position paper)

Frances C. Johnson

The Information Research Institute
Manchester Metropolitan University
f.Johnson@mmu.ac.uk

Adaptive information retrieval may use feedback to capture the context of the user's task and aspects of relevance that are hard to express in a query. However many users may find it difficult to express even the topic of interest in an initial query. The cognitive view of search is of a dynamic, problem solving activity with the user modifying the information interest as new information is retrieved. To an extent, current interactive retrieval systems support (with direct access) the query modification expected when users are engaged in understanding and developing an information interest. This paper focuses on the role and requirements of results presentation, especially summaries, in this model of interactive search and retrieval. Results presentation is an important part of a retrieval system enabling, in response, the user's query modification and system query calibration for a closer match to the relevant documents. The relevance to this workshop on adaptive retrieval lies with the interest in explicit feedback facilitating the users' evolving query. The question posed is whether there is an optimal presentation of the retrieved documents to support the process of learning and query clarification during search.

For any given task it is likely that different users will hold a different view of the information required and adopt different strategies for obtaining it. This may be partly explained with reference to Kahlthau's [1] model of information searching as a task process with various stages at which the understanding of the task changes. Each stage is characterised by a subtask and it is the associated thoughts and feelings that can influence the actions taken to advance the process. Those with greater knowledge of the information task may be at a very clear and focused stage and able to identify keywords and formulate an effective query. Those with less knowledge may be identified as being at the earlier vague and confusing stage and are more likely to browse to learn about the topic.

The key to successful search may lie in the system's ability to keep the user focused on understanding the information interest and to progress, even flow, through the stages. With this view the role of results presentation goes beyond simply indicating content of the retrieved items and hopefully why they were retrieved in response to the query. The user is further looking (from the retrieved items) for ways to conceptualise the query and use in manipulating the search. This is a very important function of the system and further research is needed to explore the optimal presentation, that is the type of information to represent content and its visualisation through understanding the interaction between results presentation and the user's search process.

Few studies have evaluated the effectiveness of different search results presentations. Dumais et al [2] found that a combination of 'clues' improved

performance and suggested that the category names help users focus in on areas of interest and the page titles help to disambiguate the category names. White et al [3] found query biased summaries were more effective than general summaries in assisting users gauge document relevance. Tombros & Sanderson [4] had similar findings and attributed this to fact that they indicated the context within which potentially ambiguous query terms were used.

The context in which the query terms appear clearly helps the user in their task and can be determined in the processing of texts. Information retrieval techniques based on term frequency distributions identify representative terms in a document for use in calculating query-document similarity and interdocument similarities for clustering. Generating document summaries are usually based on these statistical techniques, typically to extract sentences and generally to good effect [5].

Thus it is possible to present to the user summary representations indicating the key topic(s) - what the document is about – and the semantic relation held (if any) between the query terms as they appear together in the document. Furthermore, the terms with which these key terms co-occur could be shown or used to extract further sentences with a view to showing the aspects of the document/query topic(s). Representation of term distribution in the document or its structure may further indicate the meaning of the key terms in the document. It is possible to speculate that these snippets of information assist the user as they learn about the terminology and the concepts of their information interest; and, as they identify key words and formulate search expressions and tactics to manipulate the query; as well as, judging the relevance or utility of the retrieved results. Whether there is an optimal presentation, as defined above, must be addressed in further research involving users at various stages of search as they interact with results representations varying in content, size, form and structure.

Whilst this is at an early stage as a proposal it is based on decades of research on users' search interactions and research on surrogate representations. Only recently, however, has there been an interest in finding synergy between the research areas of information seeking and information retrieval. The notion that summaries serve quite specific purposes is not novel, but the proposal here aims to add a new dimension to the development and evaluation of representations specifically tailored to the users' task of formulating search. Challenging issues and questions remain for its effective implementation. These are not dissimilar to those that face the development of any information retrieval system that focuses on the users, tasks and contexts.

References

1. Kahlthau, C.C.: Seeking Meaning. Norwood, N.J: Ablex (1993)
2. Dumais, S., Cutrell, E., Chen, H.: Optimizing search by showing results in context. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Seattle (2001) 277-284
3. White, R.W., Jose, J.M., Ruthven.I.:A task-oriented study on the influencing effects of query-biased summarisation in web searching. Information Processing and Management. 39(5) (2003) 707-733
4. Tombros, A., Sanderson. M.: Advantages of query biased summaries in information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (1998) 2-10
5. Paice, C.D.: Constructing literature abstracts by computer: techniques and prospects. Information Processing and Management. 26(1) (1990) 171-186.

Queries and Relevance Assessments: The Right Context for the Right Topic

Giorgio M. Di Nunzio[1] and Nicola Ferro[1]

Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it

Abstract. We would like to discuss the problem of building a test collection for the evaluation of cross-language Information Retrieval (IR) systems. In particular, from the point of view of the experts that build the set of queries to test the performance system, and the assessors that judge the documents retrieved by the systems. Can the temporal and spatial context of a query and the user interaction history be a step forward to a more aware way to evaluate Cross-Language IR systems?

1 Introduction

The *Cross-Language Evaluation Forum (CLEF)* mainly aims at evaluating Cross-Language Information Retrieval systems that operate on multiple languages in both monolingual and cross-lingual contexts. The ad-hoc track in CLEF adopts a corpus-based, automatic scoring method for the assessment of system performance, based on ideas first introduced in the Cranfield experiments in the late 1960s. The test collection used consists of a set of "topics" describing information needs and a collection of documents to be searched to find those documents that satisfy these information needs. Evaluation of system performance is then done by judging the documents retrieved in response to a topic with respect to their relevance, and computing the recall and precision measures. The distinguishing feature of CLEF is that it applies this evaluation paradigm in a multilingual setting. This means that the criteria normally adopted to create a test collection, consisting of suitable documents, sample queries and relevance assessments, have been adapted to satisfy the particular requirements of the multilingual context.

2 The Right Context for the Right Topic

Given the experience gained being the research group responsible for the management of the CLEF technical infrastructure, we would like to bring to your attention two problems: building the set of topics, and the set of relevance judgements, in a multilingual context. In particular, would it be sensible to apply Adaptive Information Retrieval techniques for the creation of the set of queries and relevance assessments?

The creation of a set of queries suitable for a certain kind of task (ad-hoc retrieval, domain specific retrieval, geographical retrieval) is a long process. This

process requires the effort of a group of experts that have to find the right set of queries that are neither too general nor too specific; moreover, in a multilingual environment, each query should find answers also in collections of documents written in different languages and that cover different time intervals. In order to overcome this problem, the set of queries used this year in the ad-hoc track of CLEF were split into two subsets: a set of *general* queries, i.e. answers can be found in different years and different geographical locations, and a set of *specific* queries, i.e. queries that are strictly coupled with a specific collection of document and language. This fact suggests that each query has an implicit, or explicit as in this case, geographical temporal context; this context can be used to help the experts to understand whether a particular formulation of a topic is suitable or not. The idea of the context and adaptation to user behavior and experience is even more founded when you think at the process of building a query as an interactive process that requires user's feedback to an IR system in order to tune the difficulty of the query.

A similar consideration could be done for the relevance assessments. The act of judging the relevance of a subset of the documents retrieved by a system given a topic requires the assessors to scan a long list of documents. In this task human abilities and experience play an important role. The assessors of the CLEF wanted the buttons of the relevance assessment interface placed in such a way to assess as fast as possible. However, the process is so long that there is a strict limit on the number of documents that can be judged for each language. If you consider that only a few hundreds of documents are relevant over some tens of thousands, it would be vital for the assessors to rapidly focus their effort only on relevant documents. In this sense, a user interaction history, that creates the context for each particular query, may be used to skip non–relevant documents and read relevant ones only.

Solutions to these problems may be found in the use of systems like MIRACLE[1] designed for interactive Cross-Language Information Retrieval, or the use of implicit relevance feedback models like those ones presented in[2], or techniques like the Interactive Searching and Judging (ISJ) method tested by[3], or new approaches of considering the evaluation campaigns data as scientific data to be cured in order to support in-depth evaluation[4].

References

1. He, D., Oard, D.W., Wang, J., Luo, J., DemnerFushman, D., Darwish, K., Resnik, P., Khudanpur, S., Nossal, M., Subotin, M., and Leuski, A. Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi, ACM Transactions on Asian Language Information Processing (TALIP) **2**(3) (2003) 219–244
2. White, R.W., Ruthven, I., Jose, J.M., and Van Rijsbergen, C. J., Evaluating Implicit Feedback Models Using Searcher Simulations, ACM Transactions on Information Systems **23**(3) (2005) 325–361
3. Sanderson, M. and Joho, H.: Forming Test Collections with No System Pooling. Proceedings of the SIGIR 2004, (2004) 33-40 .
4. Agosti, M., Di Nunzio G.M., Ferro, N.: A Data Curation Approach to Support In-depth Multilingual Evaluation Studies, Proceedings of the Workhshop on New Directions in Multilingual Information Access (MLIA at SIGIR'06), (2006) 65–68

Adaptive Video Retrieval

Frank Hopfgartner, Robert Villa, Jana Urban

University of Glasgow
Information Retrieval Group
17 Lilybank Gardens
Glasgow G12 8QQ
United Kingdom
{hopfgarf,villar,jana}@dcs.gla.ac.uk

The Text REtrieval Conference (TREC), co-sponsored by the U.S. Department of Defence and the National Institute of Standards and Technology (NIST) supports research of information retrieval groups by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. Since 2001, a Video Track has been organised, called TRECVid. Every participant has to develop and test a multimedia retrieval system on given tasks [1], including shot boundary detection, high-level feature extraction and search.
In 2006, our team from Glasgow University participated in the search task.

For this purpose, we implemented two different video retrieval systems under the conditions of the international TRECVid workshop. The videos are segmented into shots; each shot is represented by both textual and visual features. Users can trigger retrieval cycles, browse through returned keyframes which represent video shots, play and scroll through the actual video file. For query refinement, users can give implicit and explicit relevance feedback. One objective of this work was to find out if a combination of explicit and implicit relevance feedback returns better retrieval results than a system using explicit feedback only.

1 Simulated Experiments

Both implemented systems had the same interface. For testing the objective, we ran simulated user studies using the 24 search topics of TRECVid 2005 and evaluated the results.

Another objective was to identify a model to weight existing feature categories of implicit relevance feedback. Useful categories are:

- an initial click on a keyframe (C_1)
- playing duration of a video file (C_2)
- interaction with a video such as using the pause button (C_3)

For identifying the best weighting model, we ran four simulated user studies using different values for $C_1 - C_3$ under consideration of $C_1 = C_2 = C_3$, $C_1 < C_2 < C_3$ and $C_1 > C_2 > C_3$ respectively. Again, we used the 24 TRECVid search topics from 2005 for evaluation.

2 User-based Experiments

As part of the 2006 TRECVid evaluation, we asked six users who were not familiar with our system to perform searches for a selection of the 2006 search topics. None of them was involved in the development of the system, but all had a primary degree and some an advanced degree respectively. Most of them watch TV shows on a regular basis and according to their own judgement they have a good knowledge about current affairs in general. All of them claimed to use information systems very frequently. However, they rarely use any digital video retrieval system. Each user had to work on 12 topics of the TRECVID 2006 collection and as given by the guidelines of the workshop they had a maximum time of 15 minutes for each topic. The test procedure was always organised in the same way:

- an introductory orientation session (maximum of 15 minutes)
- a pre-search questionnaire
- search session including
  - user interacting with the system (maximum of 15 minutes)
  - a post-topic questionnaire
- a post-experiment questionnaire

The total time for one session was three hours.

3 Results

Our simulated run showed that a combined system returns better retrieval results than a system supporting explicit relevance feedback only. It also showed that the system using $C_1 > C_2 > C_3$ retrieved a higher number of results than the other weight combinations.
The user-based experiments were based on the explicit relevance feedback model. One result set per topic (selected based on the number of shots marked explicitly by the user) was sent to NIST for evaluation. The results have not been published yet.

References

1. NIST: TREC: Overview. http://trec.nist.gov/overview.html (2004) last checked: 14.02.2006.

Effectiveness of additional representations in the search result presentation on the web[1]

Hideo Joho and Joemon M. Jose

Department of Computing Science
University of Glasgow
17 Lilybank Gardens, Glasgow, G12 8QQ, UK
{hideo,jj}@dcs.gla.ac.uk

1. Introduction

In the development of cognitive IR models, Ingwersen [1] discussed the importance of representing an object using multiple forms in all levels of user interactions with IR systems. The texts and document structures have been extensively exploited for effective retrieval models and search result presentations. This trend is to some extent still evident on the web. However, web pages contain a wider range of attributes than the conventional text documents, including multiple colours, layouts, and images. These visual elements of documents are likely to have an effect in the search process [2], thus, they can be exploited as a component of surrogates. Therefore, we carried out a user study to compare the effectiveness of textual and visual features as *additional representations* in the search result presentation. Unlike existing work [3], in our experiment, participants were involved in all aspects of searches.

2. Summary of experiment

We devised four layouts of search result in the experiment. Layout 1 was based on Google's result. Layout 2 had additional textual representation based on the top ranking sentences (TRS) [4]. Layout 3 had additional visual representation based on a thumbnail image of web pages. Finally, Layout 4 (shown in Fig. 1) had both the TRS and thumbnail in the presentation.

Fig. 1. Search result with TRS and thumbnail (Layout 4)

Twenty-four participants (6 females, 18 males) were recruited for the experiment. Each participant carried out four search tasks using a different order of the four lay-

[1] This work was supported by EPSRC (Grant ref: EP/C004108/1).

outs. The search tasks used in the evaluation were: 1) background search task, 2) decision making task, 3) known item search task, and 4) topic distillation task.

We did not find a significant difference among the layouts regarding the time completion time. However, when an additional representation was available in the interface (i.e., Layout 2 to 4), participants appeared to submit more queries using a wider range of words, compared to Layout 1. Therefore, there seems to be a relation between the level of document representation and user's query re/formulation process. We speculate that an increased level of document representation can facilitate user's query re/formulation process. As for the browsing of search results, the number of click-through URLs was fewer in Layout 2 to 4 compared to Layout 1. Participants were also viewing more search results in Layout 2 to 4 than Layout 1. Therefore, prticipants appeared to make a judgement of retrieved documents on the search result more frequently, compared to Layout 1. This suggests that an increased level of document representation also has an effect on user's browsing process.

However, the effectiveness of TRS and thumbnail was often inconsistent across the search tasks. This was partly found in participants' perception on the usefulness of laytout features (shown in Table 1). As can be seen, TRS was significantly correlated with other textual representations such as title and Google snippet, while the thumbnail had a significant negative correlation with the snippet, and positive correlation with the URL. This suggests that the effectiveness of the textual and visual additional representations can be mutually exclusive. Therefore, this study calls for further research on the understanding of users' search contexts and adaptive technique to capture their needs in an appropriate context. Also, this study implies that a careful consideration might be required for the selection of additional representation when the level of document representation is to be increased in interface design.

Table 1. Correlation of layout features contribution (Spearman's coefficient; N=48)

	Title	**Snippet**	**TRS**	**Thumb.**	**URL**	**Size**	**Type**
TRS	**.410**	**.314**	1.000	-.175	-.202	.010	.147
Thumb.	.210	**-.265**	-.175	1.000	**.284**	**.247**	.051

3. References

1. Ingwersen, P., *Information Retrieval Interaction*. 1992, London: Taylor Graham Publishing.
2. Tombros, A., I. Ruthven, and J.M. Jose, *How Users Assess Web Pages for Information Seeking*. Journal of the American Society for Information Science and Technology, 2005. **56**(4): p. 327-344.
3. Dziadosz, S. and R. Chandrasekar. *Do thumbnail previews help users make better relevance decisions about web search results?* In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 2002. Tampere, Finland: ACM.
4. White, R., J.M. Jose, and I. Ruthven, *Using Top-Ranking Sentences to Facilitate Effective Information Access*. Journal of the American Society for Information Science and Technology, 2005. **56**(10): p. 1113-1125.

Using local contexts to slice and dice the search results on the web*

Hideo Joho and Joemon M. Jose

Department of Computing Science
University of Glasgow
17 Lilybank Gardens, Glasgow, G12 8QQ, UK
hideo,jj@dcs.gla.ac.uk

1. Introduction

An effective way to group retrieved documents has been an important issue in Interactive Information Retrieval (IIR). A recent study suggests that faceted grouping can be a promising alternative to clustering techniques [1]. However, the success of faceted grouping seems to rely on sufficient knowledge of collection structure. In this paper, we propose an alternative approach to faceted search and browsing based on the local contexts of query terms. We define the local contexts as the words that appear in the surrogate of search results. The use of local contexts is appealing since it requires less knowledge of the collection than existing approaches. The proposed interface offers an area called *Workspace* where searchers can explore the search result without interfering the original result (See Fig. 1).

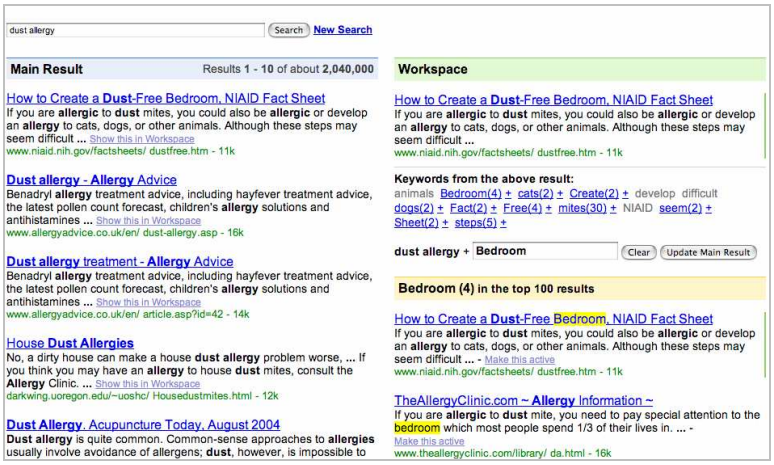


Fig. 1. Screenshot of proposed interface

The Workspace area (Right) is activated by clicking one of the records in the Main Result area (Left). The keywords in the document surrogate of clicked documents are presented to a searcher where they are used as pseudo-facets to explore the search results. Alternatively, the searcher can type any words in the workspace keyword box to retrieve a subset of retrieved records. As can be seen, since all pseudo-facets are extracted from the surrogate of retrieved documents, our approach assumes little about the collection structure for the implementation.

2. Summary of experiment

Twenty-four participants (2 females, 22 males) were recruited for the experiment. Each participant carried out four search tasks using a different order of two interfaces: Baseline (Main Result area only) and Workspace. The topics used in our experiment are 1) Dust allergy in workplace; 2) Music piracy on the Internet; 3) Petrol price; and finally, 4) Art galleries and museums in Rome.

As an overall assessment of the interfaces, participants were asked to indicate the preference of two interfaces based on their experience of four search tasks at the end of experiment. 21 out of 24 (87.5%) indicated that they preferred the Workspace interface over the Baseline interface. Participants' subjective assessment on the usefulness of the Workspace interface was significantly better than the Baseline interface. This suggests that participants welcomed the functionality offered by the workspace. Our results also indicate that participants' motivation to access the workspace differs over the complexity of search tasks. With a lower complexity task, participants typed their own words in the workspace keyword box, while they tended to select the extracted pseudo-facets more frequently in a higher complexity task.

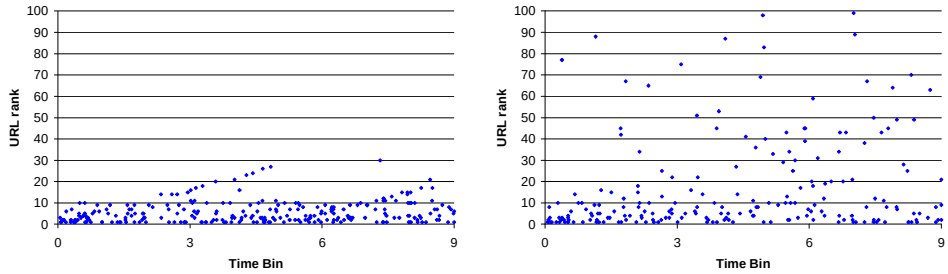


Fig. 2. Rank position of bookmarked pages: Baseline (Left) and Workspace (Right)

Figure 2 shows the rank distribution of the pages bookmarked by participants during the experiment. As can be see, while most bookmarked pages were ranked within the top 30 in the Baseline, the distribution was scattered in a wider position in the Workspace interface. This suggests that participants were exploring the search results and finding relevant information regardless of the original ranking in the Workspace interface. A problem of the current implementation is sometimes it presents a limited range of pseudo-facets to the searchers. We are currently investigating the effects of query-biased sentences [2] in populating the pseudo-facets.

3. References

1. Yee, K.-P., et al. (2003) Faceted metadata for image search and browsing. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2003: ACM.

2. Tombros, A. and M. Sanderson. *Advantages of query-biased summaries in information retrieval*. in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1998. Melbourne, Australia: ACM.

* This work was supported by EPSRC (Grant ref: EP/C004108/1).

A sentence-based ostensive browsing and searching on the web*

Robert D. Birbeck, Hideo Joho, and Joemon M. Jose

Department of Computing Science
University of Glasgow
17 Lilybank Gardens, Glasgow, G12 8QQ, UK
birbeckrd,hideo,jj@dcs.gla.ac.uk

1. Introduction

The ostensive model assumes that a user's information need is in nature dynamic and developing, thus, a recently accessed object can be seen as more indicative to the current information need than previously accessed ones [1]. The model has been mainly applied to image retrieval [2, 3]. In this study, we applied the ostensive model to the web retrieval by using the top ranking sentences (TRS) [4] as a means of browsing search results as well as capturing relevance feedback implicitly.

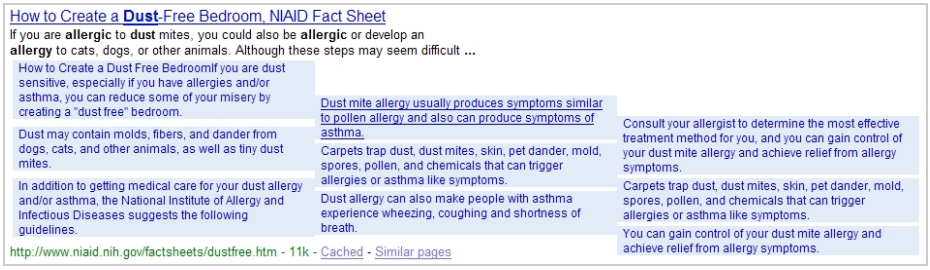


Fig. 1. A screenshot of proposed ostensive browsing interface

Figure 1 illustrates the design of our ostensive browsing interface. Firstly, search results were supplemented by up to three TRS in the interface. When a user hovered the mouse pointer on a sentence, three new TRS were presented to the user. The candidate sentences were taken from the top 30 URLs. They were ranked by an ostensive model function which gave a higher weight to the words that appeared in more recently accessed sentences. The interface also had a term suggestion feature which expand an existing query with the words that had the highest score.

2. Summary of experiment

Twenty-four participants were recruited for the experiment. Each participant carried out three search tasks using a different order of the three interfaces: Baseline (Static TRS only), Ostensive 1 (Ostensive TRS browsing), Ostensive 2 (Ostensive TRS browsing + Term suggestion). While all interfaces presented 10 URLs per result page, participants had an access to the top 30 URLs using TRS in the Ostensive 1 and 2 interfaces. The search tasks used in the evaluation were: background search task, de-

cision making task, and finally, many items task. Participants were asked to bookmark the pages when perceived relevant information was found.

As an overall subjective assessment of the interfaces, participants were asked to indicate their preference of the interfaces at the end of experiment. The result shows that participants significantly preferred Ostensive 1 and 2 to Baseline. Other subjective measures suggest that participants often found Ostensive 1/2 easier to browse the search results and find relevant information compared to Baseline. They also tended to find Ostensive 2 easier to re/formulate queries during the tasks. While the difference was not statistically significant, participants used a wider range of words in Ostensive 2 compared to Baseline and Ostensive 1. These results suggest that the ostensive presentation of TRS had a positive effect on participants' browsing of search results and query re/formulation process. On the other hand, the time taken to complete the tasks and number of bookmarked pages were comparable across the interfaces.

	Title	Snippet	TRS	URL	Size	File Type
Baseline	2.0 (1.0)	2.5 (1.0)	3.0 (1.7)	4.5 (2.3)	6.0 (1.6)	5.4 (1.7)
Ostensive 1	1.8 (0.9)	2.0 (1.2)	1.9 (1.0)	4.2 (2.1)	6.0 (1.7)	5.8 (1.6)
Ostensive 2	1.6 (0.9)	1.8 (1.0)	1.6 (0.8)	4.5 (2.1)	6.1 (1.6)	6.0 (1.6)

Table 1. Contribution of layout features (Range: 1-7; Lower = Stronger)

An interesting observation was that participants tended to rate the contribution of TRS in relevance assessments higher in Ostensive 1/2 interfaces compared to Baseline (See Table 1). When the ostensive presentation of TRS was available, participants also tended to rate the contribution of title and snippet higher than Baseline. These results indicate that the ostensive presentation can lead to an increased level of awareness of TRS as well as other components of document surrogate in the search results.

The experiment also suggests that the current presentation of suggested terms should be improved. In the current implementation, the top six words were appended to an existing query as participants accessed to TRS. However, participants tended to accept all words or delete them all before submitting a new query. A more effective way to select words from suggested terms should be devised and this is one of our future work.

3. References

1. Campbell, I. and C.J. van Rijsbergen (1996) The Ostensive Model of Developing Information Needs, In *Proceedings of CoLIS2*, p. 251-268.
2. Campbell, I. (2000) Interactive Evaluation of the Ostensive Model Using a New Test Collection of Images with Multiple Relevance Assessments. *Journal of Information Retrieval*, **2**(1): p. 89-114.
3. Urban, J., J.M. Jose, and C.J. Van Rijsbergen (2003). An Adaptive Approach Towards Content-Based Image Retrieval, In *Proceedings of CBMI'03*, p. 119-126.
4. White, R., J.M. Jose, and I. Ruthven (2005). Using Top-Ranking Sentences to Facilitate Effective Information Access. *JASIST*, **56**(10): p. 1113-1125.

* This work was supported by EPSRC (Grant ref: EP/C004108/1).

Identifying Features for Relevance Web Pages Prediction *

A. Masegosa †, H. Joho ‡ and J. Jose ‡

andrew@decsai.ugr.es, {hideo,jj}@dcs.gla.ac.uk
† Department of Computer Science and A.I., University of Granada
‡ Department of Computer Science, University of Glasgow

1 Introduction

In a web search task, we consider a web page relevant when it contains what we are looking for. The text content of a web page has been widely used to assess the relevance of web pages, but there are several studies [1, 2] that show the existence of other factors involved in the relevance assessment related with the structure, non-textual items, etc. of web pages.

In this work we analyze more than 150 web page features in order to investigate, using a machine learning approach, which ones are the most informative about the relevance of web pages.

2 Feature Description and Experimental Results

We have extracted 150 features and grouped them in the following sets:

Textual Features (14 Features): Features respect to the textual content of a web page are evaluated. : Number of Words, Entropy of the word distribution, number of words in anchor text..

Visual/Layout Features (71 Features): Features respect to the visual and layout content or appearance of a web page: the height of the document, the height mean of the images... Another subset is related with the measure of the number of occurrences of the *center* tag, *p* tag... and special attributes of these tags like *style*, *size*...

Structural Features (18 Features): Features related with structural aspect of a web page like the page rank, the number of links...

Other Features (47 Features): Special features like the number tags related with event management of a web page and special tags like script tag, onclick attribute...

* This work has been supported by ALGRA project (TIN2004-06204-C03-02), FPU scholarship (AP2004-4678) and EPSRC (Ref: EP/C004108/1)

For this analysis, we used the experimental data collected from a previous study [3] where 24 users were asked to perform different web search tasks and indicated the relevance of web pages. From the data, we obtained 737 unique click-through web pages, of which 362 were non-relevance and 375 were relevance.

In this table we show for each one of the analyzed features subsets the number of selected features (Num.), their predictive accuracy and we list the five most important features in each one of the features sets:

Feature Subset	**Num.**	**Accuracy (%)**	**Best 5 Features**
Textual	9	55.73 v	entropy percentage of anchor and document text, **num. digits**, disk size, number of upper words
Visual Content	12	53.99	percentage of area background images, image disk size, num. link style, mean of width images, **num. images**
Layout Html Tags	13	53.57	new line, paragraph, **meta**, **bold**, division
Visual Html Tag Attributes	12	56.65 v	**alternate text** (alt), **border of a table**, style, size, alignment
Structural	11	51.75	num. of outside/inside links, num. of html links, page rank host page, number of url levels
Other features	8	52.99	src attr., input tag, script tag, value attr., onclick attr.

v statistically significant improvement respect to the random assignation

Our feature lists appear to support the results of existing studies. For example, Thombros et al. [1] suggested the presence of digits and table data and Fox et al. [2] the number of images as indicators of relevance. And although there is limited known features used in search engines, the presence of the query terms in meta data, bold words and alternate text has been suggested as important factors for ranking and these tags have been also selected in our work. Therefore, we speculate that the extraction of other features found in our analysis have also a potential effect to improve the accuracy of relevance estimation in adaptive IR systems. In this study we investigated features independently in each group. In the future, we are interested in investigating the integrated set based on

References

1. Tombros, A. et al.: How users assess web pages for information seeking. Journ. American Society for Information Science and Technology **56**(4) (2004) 327–344
2. Fox, S. et al.: Evaluating implicit measures to improve web search. ACM Trans. Inf. Syst. **23**(2) (2005) 147–168
3. Joho, H., Jose, J.M.: A comparative study of the effectiveness of search result presentation on the web. In: Advances in Information Retrieval, 28th European Conference on Information Retrieval, LNCS (2006) 302–313

A System for Adaptive Information Retrieval

Ioannis Psarras[1], Joemon M. Jose[1]

[1] Department of Computing Science, University of Glasgow, Glasgow, G12 8QQ

1 Introduction

Recent studies, such as [2] and [3], have demonstrated the shortcoming of modern search engines, by highlighting that such tools fall short in organizing and managing user information needs.

Often such information requirements change by sliding into new topics. The only way to satisfy such needs is to search on a continuous basis, that is keep looking for information regularly. On the other hand, similar to changes in user interests/needs, documents on the web keep changing as well. Unfortunately, no search engines currently help the user in finding documents with respect to their dynamic information needs.

In this paper, we describe the design and evaluation of a personal information assistant aiming to profile the volatile requirements of users and present new information with respect to their needs. Our system, called PIA (Personal Information Assistant), is able to adapt to the changing needs of users, manage the multiple facets of user profiles, and pro-actively fetch and recommend additional documents on a regular basis.

We have evaluated the system using a task-oriented evaluation methodology, where nineteen users used the system regularly for 7-10 days. A direct comparison with the most successful commercial search engine (Google) was made to observe the performance of our system against Google, a very effective information retrieval tool. The evaluation results illustrate that the Personalized Information Assistant is effective in capturing and satisfying users' evolving information needs and providing additional information on their behalf.

2 System Overview

Personal Information Assistant was developed as an adjunct to the current web search engines. The main interface features a profile editor, to allow users to bypass the implicit profiling process and amend their interests explicitly, as well as a search engine, which forwards all queries to Google. The results are parsed and presented to the browser. At this stage, the user's profile gets updated to exploit the information gathered from the previously issued search. At some point in the future, the assistant will analyze the user's profile and attempt to retrieve additional relevant documents regarding the user's evolving needs.

3 The Profiling Algorithm

Past solutions, like [1, 4], integrated profiling algorithms and represented users' profile as a single weighted keyword vector. However, user interests are multiple and

must be represented accordingly in a person's profile. PIA recognises the multiple aspects of users' profiles and represents them as separate weighted keyword vectors.

The profiling algorithm, integrated in our system, starts with retrieving a set of representative terms, by continuously monitoring user interaction and exploiting implicit user interest in documents, during each search iteration. Having extracted a set of terms from visited documents during the recent search, a new interest is created or an existing one is amended to take into account the retrieved keywords. We used clustering techniques to detect various facets of users' interests and to decide whether a set of terms should be translated as a new user interest or as part of an existing interest.

This profiling strategy takes place in each search iteration in order to allow the system to adapt to user changing needs. Two interests can be merged together, in case their vectors are adequately similar, while new interests can be created during this process. At the end of the profiling process, user interests have been updated to adapt to the new information gathered during the recent search.

4 The Recommendation Process

At regular time intervals, the system will read and analyze user profiles and formulate a new query based on the keywords of each user interest. The query terms are chosen by extracting a number, between five and eight, of frequent words from an interest. A new search is issued, using the formulated query terms, and a number of the top ranked documents are recommended to the user.

5 Conclusion

We have designed, deployed and evaluated a system aiming to supply users with up-to-date information regarding their personal needs. By using an implicit feedback gathering model, we eliminate the necessity of forcing users to create their profiles explicitly. By formulating queries based on the users' interests and automatically seek more information on the web, the assistant recommends additional documents that might be of interest to the users.

References

1. Chen, L., & Syraca, K., Webmate: A Personal Agent for Browsing and Searching, Proceedings of the 2nd International Conference on Autonomous Agents, 132-139.
2. Jansen, B.J. and Pooch, U. (2000). A Review of Web Searching Studies and a Framework for Future Research. Journal of the American Society for Information Science and Technology. 52(3), 235-246.
3. Jansen, B.J., Spink, A. and Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of users on the Web. Information Processing and Management. 36(2), 207-227.
4. Lieberman, H., Dyke, N. W. V. and Vivacqua, A. S. Let's browse: a collaborative Web browsing agent. In Proceedings of the 1999 International Conference on Intelligent User Interfaces (IUI'99), pages 65–68, Los Angeles, CA, USA, 1999. ACM Press.

Combining Image Organisation and Retrieval to Overcome the Semantic Gap in CBIR

Jana Urban

Department of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK
jana@dcs.gla.ac.uk

1 Introduction

Content-based image retrieval (CBIR) is an intrinsically hard problem. Manual labelling is impractical for most purposes and automatically extracted content-based features do not describe what humans recognise and associate with an image, referred to as the semantic gap. The semantic gap complicates the query formulation process for the searcher. Moreover, image meaning is subjective and context-dependent. Finally, information needs are often vague and dynamic, since image searches are usually coupled with creative tasks. These problems render current image retrieval systems difficult to use. Unlike most previous work in the field which has studied the retrieval system as a self-contained problem, the approach described here takes a holistic view, in which information access is considered as part of a larger work process. By taking into account the design of both the interface and retrieval algorithms, all of these intrinsic issues of image retrieval are addressed together.

The starting point in creating a more user-friendly system was to redesign the interaction process between user and system. Based on analysing user studies [1, 2] the organisation of information has been found to help structure the thought process of the searcher. Therefore, the proposed system, *EGO* (Effective Group Organisation), combines image management and search. This is achieved by incorporating a workspace in the interface, allowing the user to organise search results into groups on the workspace. A recommendation system, which suggests new images for existing groups, assists the user in this task. The grouping process is incremental and dynamic: through usage a semantic organisation emerges that reflects the user's mental model and their work tasks. Hence, *EGO* aims to represent the context in which the images are used.

The usage information, in the form of relationships between images grouped together, is further used as a semantic feature in the proposed retrieval model [3]. In addition to these semantic relationships, visual and textual features are modelled in a single graph, which uses the theory of random walks as the basis for the retrieval algorithm implemented on the graph.

2 Evaluation

The benefits of *EGO* were studied in a user-centred, task-oriented evaluation involving 24 participants and 6 design-oriented image search tasks with different

types of information seeking scenarios [4]. *EGO* was compared to a traditional relevance feedback interface. The evaluation hypotheses were: (1) *EGO* leads to an increased effectiveness and user satisfaction; (2) it helps to conceptualise and diversify tasks; and (3) it helps to overcome the query formulation problem.

The participants preferred the proposed approach and the perceived effectiveness was better. While the relevance feedback system increased performance for narrow search tasks, the required effort to complete more open, design-oriented tasks was lower with *EGO*. Users indicated that *EGO* helped to analyse and explore their tasks better. The resulting groups people created on the workspace reflected task complexity. The users also had more problems with the relevance feedback facility than with *EGO*'s recommendations, although the underlying retrieval system was the same. In the recommendations they could see which images contributed to the query, while at the same time hiding the details of the retrieval mechanism, hence alleviating the query formulation problem.

3 Conclusion

The grouping process allows the user to organise search results based on semantic concepts. The system then adapts its internal image representation to reflect these concepts. Both these factors help to encode the intended meaning of the images. Further, the query formulation problem is mitigated, since groups are considered as implicit search requests. Finally, groups emerge as facets of the user's information need, helping the searcher to conceptualise and develop complex and dynamic needs. Moreover, the system is informed of changes in information need when the user switches back and forth between groups. Altogether, *EGO* creates an environment, in which the meaning of an image is interactively defined, the query formulation problem is mitigated, and time-varying information needs can be expressed. Hence, it is a user-centred approach that comes close to bridging the semantic gap.

Acknowledgement This work was supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content—K-Space—and by the EPSRC (Grant ref: EP/C004108/1). This publication only reflects the authors' views.

References

1. Nakakoji, K., Yamamoto, Y., Takada, S., Reeves, B.N.: Two-dimensional spatial positioning as a means for reflection in design. In: Proc. of the Conf. on Designing Interactive Systems (DIS'00), New York, NY, USA, ACM Press (2000) 145–154
2. Rodden, K.: How do people organise their photographs? In: Proc. of the 21st BCS IRSG Colloquium on IR, Electronic Workshops in Computing. (1999)
3. Urban, J., Jose, J.M.: Adaptive image retrieval using a graph model for semantic feature integration. In: Proc. of the 8th ACM SIGMM Int. Workshop on Multimedia Information Retrieval (MIR'06), ACM Press (2006)
4. Urban, J., Jose, J.M.: Evaluating a workspace's usefulness for image retrieval. ACM Multimedia Sytems Journal (Special Issue on User-Centered Multimedia) (2006)

A Formal Approach to Information Retrieval (Search) based on Quantum Mechanics

Sachi Arafat

University of Glasgow
Glasgow, UK
sachi@dcs.gla.ac.uk

Abstract. The main issues in traditional retrieval are user, data and relevance modeling; and deducing of retrieval strategies, user-interaction/interface and search context. Current research on relating all these aspects is mostly ad-hoc. The evaluation methodologies to judge between effectiveness of the above components are weak in that they are specific to the corpora or the user-experimentations on which they are highly dependent. It then becomes difficult and expensive to compare/contrast systems, especially interactive IR systems. These problems are due to the inherent weaknesses in the modeling apparatus, preventing adequate capture of the entire search process. Our research goal is to formally unify all these aspects under one framework to simplify comparison between search systems and their evaluation. For reasons outlined here we use the modeling apparatus of quantum theory as a means to achieve this goal. What we found (and surprisingly so) through developing our framework is that the *semantics* and operational methods of quantum mechanics (QM) are crucially more relevant in understanding IR than the mathematical formalism of QM is in unifying some already existent formal models.

Program

09:00-09:45	*Invited talk*	
	Nick Belkin	*Getting Personal: Personalization of Support for Interaction with Information*
09:45-10:45	*Provocative Position papers (30 minutes each)*	
	Kalervo Järvelin	*Simulating Searcher's Feedback Quality and Effort in Interactive IR*
	Joemon Jose	*Issues in the Research on Adaptive Search Systems*
10:45-11:30	Coffee Break & Poster Presentations	
11:30-12:15	*Invited talk*	
	Noriko Kando	*A Model of IR Testing and Evaluation: From Laboratory towards User-Involved*
12:15-13:15	*Provocative Position papers (30 minutes each)*	
	David Harper	*Employing User Relevance Assessments for Measuring Retrieval Effectiveness*
	Mark Sanderson	*Test collections for all*
13:15-14:00	Lunch Break & Poster Presentations	
14:00-14:45	*Invited talk*	
	Ellen Voorhees	*Building Test Collections for Adaptive Information Retrieval: What to Abstract for What Cost?*
14:45-15:45	*Discussion Groups (2-3 Groups)*	
15:45-16:15	Coffee Break & Poster Presentations	
16:15-17:00	*Report Back/Recommendations*	
17:00-18:00	*Panel on Adaptive Retrieval and Evaluation*	
18:00-xx:xx	Pub Crawl	