Proceedings of the

Glasgow –Strathclyde

Information Retrieval Workshop

Funded by Synergy

22nd of October 2004
IEEE Teachers Building
St Enoch's Square, Glasgow

Organisers:

Leif Azzopardi	leif@dcs.gla.ac.uk
Ian Ruthven	ian.ruthven@cis.strath.ac.uk
Fabio Crestani	fabio.crestani@cis.strath.ac.uk
Keith van Rijsbergen	keith@dcs.gla.acuk

Table of Contents

Introduction

The Information Retrieval Workshop was held on the 22nd of October, 2004 at the IEEE Teachers Building in St Enoch's Centre, Glasgow. The event attracted over twenty Information Retrieval Researchers from the University of Glasgow and the University of Strathclyde, comprising of Academic Staff, Research Assistants and PhD Students. The attendees included: Ida Abdullah, Sachi Arafat, Azreen Azman, Leif Azzopardi, Mark Ballie, Di Cai, Mathew Chalmers, Fabio Crestani, Heather Du, David Elsweiler, Magdah Gharieb, Forbes Gibb, Mark Girolami, Claudi Hauff, Srikant Jakilinki, Iraklis Kampanos, Monica Landoni, Christina Lioma, Emma Nicol, Iadh Ounis, Vassillis Plachouras, Fabio Simoeni, Abhishek Sharma, Simon Sweeney, Jana Urban, Keith van Rijsbergen, Reede Ren, Ian Ruthven and Huang Zheng.

During the course of the workshop 19 presentations were given by various members of the IR groups; nine from Glasgow and ten from Strathclyde. The presentations provided an opportunity for participants to share their current research interests with the local IR community and to see where commonilaties existed to develop collaborative research endeavours. Delivery of the presentations was restricted to only five minutes and coupled with discussion panels to elaborate on the themes of the session.

A by product of the format was interactive and engaging sessions which resulted in a greater awareness of the current research from which future work could be established by utilising the skills and knowledge within the local community. These proceedings contain the main content presented during the workshop and serve as a brief overview of Information Retrieval research in the Glasgow area. Several potential collaborations are being investigated as a result of this workshop and another similar event between the two groups is going to be organised for next year.

After the workshop two prizes were awarded for the best presentation. These were awarded to David Elsweiler and Mark Ballie. Congratulations!

The running of the workshop would not have been possible without the funding from Synergy, the organisational assistance from Jacqui Brannan and editorial assistance from my sister, Cindy Azzopardi. Thank you.

Best Regards,
Leif Azzopardi

Resources for identifying noun-phrases

Noorhidawti Abdullah

noorhidawati.abdullah@cis.strath.ac.uk

Areas of Interest

Automatic classification and indexing, summarisation, e-books, natural language processing, digital library

Description of Current Research

Currently I am doing research on automatic back-of-eBook indexing. Readers of digital documents or eBooks are normally provided with three main methods for locating relevant sections of text: the table of contents (ToC), which defines the logical structure of the book; a full text search facility; and bookmarks. However, research (Landoni and Gibb, 2000; Landoni, Wilson and Gibb, 2000; Landoni, Crestani and Melucci, 2000; Landoni, Wilson and Gibb, 2001; Wilson, Landoni and Gibb, 2002) has highlighted that one key functionality which eBook readers expect is rarely provided within an eBook environment: the back-of-the-book index (BoBI). The value of a BoBI lies in the fact that it provides a structured and contextualised list of major concepts which are contained in the text and which can be used to facilitate access to relevant sections of a document. Therefore, the purpose of this research is to develop and test methods for automatically generating BoBIs for digital documents or e-content (for scientific and technical documents) in order to improve the quality of retrieval and the speed of access to relevant sections of text. Only limited research has taken place with the respect to the generation of BoBIs and it has been restricted to the extraction of noun phrases supported by statistical analysis. The novelty of the approach in this research lies in the use of more complex linguistic units and the integration of NLP, statistical indexing design and lexical resources. The tool should be of use to authors, publishers and readers of eBooks and will have wider application within the digital library context.

Project Proposal

Natural language parsing (NLP) has been widely investigated as an enhancer of information retrieval. NLP tools can, as a minimum, produce a candidate list of single terms (which we refer to as atomic concepts) based on morphological features (e.g. nouns, adjectives and verbs) which represent entities, their attributes and operations performed upon them or by them. Further, the relationships between these elements can be elicited using morpho-syntactic information (e.g. subject, object and main verb) to create larger units. Within the automated indexing community the noun phrase has been consistently proposed as the most important text unit for generating compound descriptors for a text (Evans and Zhai, 1996; Haddad, 2002). The ability to identify and uniquely interpret noun phrases within documents should, it has been argued, improve both precision and recall by preserving context, recognising and conflating different surface forms, and introducing the possibility of synonym control (Gay and Croft, 1990)

Not all noun phrases may be useful descriptors for a text, however: they may require filtering or further processing in order to generate suitable subject indicators. In addition, although noun phrases have been widely used there are other linguistic units and features which can be applied (Rinaldi et al, 2002; Schwitter, Molla and Hess, 2000) though these have not been as fully explored. Other conventions which will be explored include: verbal cues, such as 'in conclusion', 'to summarise' etc., and classes of noun phrases. For example, an analysis of medical texts has shown that the generally used technique for introducing a topic is the indefinite noun phrase (Hein, 1989). Subsequent references to a topic are then made through definite anaphoric noun phrases. These anaphoric noun phrases are also important clues to what Hein refers to as the background profile of area of discourse; that is, they indicate domain-specific knowledge which the reader must possess in order to comprehend the text. In addition the use of other features such as prepositional attachments can be used to identify relationships which exist between noun phrases (to create what we refer to as macromolecular concepts). Statistical information on the behaviour of concepts should also help to gauge their relative importance to a section or sub-section of text. A key feature of this study will be to combine a range of knowledge resources to improve the selection of noun phrases and to increase the scope to larger linguistic units.

Simulated interactive evaluation toolkit

Sachi Arafat

sachi@dcs.gla.ac.uk

Areas of Interest

SIMINEV (SIMulated INteractive EValuation) Formal Methods/Models, Cognition Modelling, Quantum Modelling, Simulation

Description of Current Research

Human-human information retrieval (HHIR) is the optimum of the imitation that is human-computer retrieval. From this perspective any IR system is simulating a cognitive agent. The richer of the simulations are those found in interactive retrieval where the language of interactions approximates speech, facial expressions, gestures and other communication protocols & methods in the HHIR domain.

In order to 'better' simulate the cognitive agent one requires to model it and understand how it responds to changes that occur as a HHIR session proceeds. In my work I model the cognitive agent as a system that undergoes a set of state changes during a retrieval session. We find that describing such changes using the principles and mathematics of quantum mechanics (QM) is adequate as it accounts for many cognitive phenomena. Thus far we have such a mathematical framework that formalizes the state changes that occur in the simulated cognitive agent (retrieval system) during a retrieval process.

In this framework we can formulate retrieval strategies that agree with our cognitive model. Currently I am trying to evaluate the claim that such strategies are (a) natural with respect to cognition (b) effective for retrieval. The baseline is Campbell's ostensive retrieval system. At the moment we use this system replacing the recommending methods by our strategies. In summary my research aims to investigate and formalize ostensive retrieval with respect to HHIR.
Project Proposal: In order to test effectiveness of retrieval strategies we have defined some measures for ostensive retrieval quite closely tied to the path model in Campbell's interface. The nature of the measure coupled with the way the interface updates per interaction suggests that evaluation of the strategies by simulating user interaction is more adequate in this context than user testing. The Lemur toolkit was used to develop a simulation to evaluate a retrieval strategy generated by the framework: the SIMINEV (SIMulated INteractive EValuation) project. Using a certain measure which can be used to rank search paths based on a semantic, much insight was provided as to the nature and effectiveness of the retrieval strategy, especially since the strategy is derived from cognitive principles.

As research proceeds new strategies will result meaning further simulations will have to be devised to evaluate them. There is a more natural link between our approach to IR and general user simulation since any new strategy for the system that imitates a cognitive agent can also be used to simulate the user in the evaluation phase. Hence as new strategies for the system are formulated a dual set of heuristics for user simulation are automatically suggested.
Retrieval strategies in this context have a broader definition encompassing underlying ranking/recommending methods as well as interface changes. This allows consideration of a generic set of interfaces and interactions by the strategies and lets us simulate retrieval with them.
Researchers at Glasgow have already used simulation of interactive retrieval for evaluation purposes which has been especially detailed in White et al 2004. In that paper a system accepting relatively complex interactions is evaluated by simulation. Such interactive systems have been shown to be effective and as they are further developed the software for their evaluation by simulation must similarly be updated.

The SIMINEV project is thereby proposed as an open source simulation toolkit that could be developed in collaboration to combine current approaches to user simulation for evaluation of interactive systems. Main outcomes foreseen are:

- An add-on to the Lemur Toolkit to provide researchers with common simulation functionality for evaluating interactive systems
- A study to determine the properties of interactive systems that encourage or discourage evaluation by simulation
- A study to investigate the retrieval scenarios in which combining simulation with user-based evaluation would be effective.
- Comprehensive set of tools and guidelines aimed at minimizing time and money spent on evaluation of interactive retrieval systems.

The initial outcome requires the creation of an especially non-trivial, extensible software architecture which must accommodate for high variability in several simulation categories including different user models, interfaces, retrieval strategies and interaction methods. In doing this it is likely that issues regarding limitations to simulation may surface meaning that in some scenarios a combination of both simulation and user testing are required (outcomes two and three). At least four people have shown interest in such a project at Glasgow and I hope others will join in. I expect much of it can be developed separately according to the interests of the collaborators once a rough idea of the architecture surfaces.

Position Statement

Azreen Azman

azreen@dcs.gla.ac.uk

Areas of Interest

Web recommendation, Adaptive websites, Data mining for information retrieval, Adaptive information retrieval, Collaborative filtering and user interaction and modelling

Description of Current Research

In the Web environment, information and users are heterogeneous in nature. Web contains information for number of different subjects, geographical areas, types, formats and so on. The context of user's query plays a vital role in the retrieval performance of a search engine.

Furthermore, user approaches an IR system when he/she requires some information, be it a query-based system (e.g. search engines) or browse-based system (e.g. web sites). His or her information need is dynamic, which means that he or she has different need at a different time. In addition, user's information need is multi-dimensional, whereby different users has different needs. Therefore, a good IR system should be robust enough to correctly identify context of user's information need.

The main motivation of my research is to answer a very simple question; whether or not an IR system is adaptive enough to identify context of user's information need? I am interested in the adaptive approaches to infer information need of a user.

The problem of inferring user's information need can be tackled by observing user's behaviour when interacting with the system, such as selecting document from the ranked list or while browsing. The main assumption is that user's selection or browsing strategy is guided by his/her information need. I am investigating this problem in the area of Web recommendation system.

Evaluating interactive information retrieval in simulated and real environment

Leif Azzopardi

leif@dcs.gla.ac.uk

Areas of Interest

Language modelling, summarization, simulated interactive retrieval, probabilistic models, contextual information retrieval

Description of Current Research

My Phd work has been examining ways in which we can incorporate contextual evidence in the retrieval process using the language modelling approach (Ponte and Croft, 1998). Specifically, I have been doing this by ingraining a user bias when modelling documents, i.e. creating a document representation with respect to the user's understanding of the collection - and this actually instantiates the Cluster Hypothesis within the Language Modelling framework. Other aspects of Language Modelling I have investigated are: the underlying assumptions of the model, where the a document's relevance is assumed to correlated with the document producing a query; and I have attempted to address the theoretical problems within the framework by proposing alternative models for relevance feedback.

Project Proposal

The typical view of an Information Retrieval System (IRS) has been that a system takes as input a set of documents and a query (which represents the user's information need) and returns a set of documents. In response to the set of returned documents the user engages in feedback which is entered back into the system (van Rijsbergen, 1979ir). In order to evaluate an IRS, standard methods of interactive have been defined (such as ad-hoc querying, query reformulation, query expansion, relevance (and pseudo) relevance feedback, etc). Each interaction is typically evaluated or assessed according to a specific simulation of that interaction. This results in a very clinical and disjunct view of the interactive process, as there is no consideration to the combinations of interactions to satisfy ones information needs (multiple queries, examining different documents, past interaction, etc). Hence a major challenge is in evaluating Interactive Information Retrieval which goes beyond just one form of interaction and one stage of interaction (i.e beyond, enter query, get results, give relevance feedback, get new results, evaluate.) To examine more complex interactions would usually involves running user evaluations. However, recently White et. al. (2004) proposed a simulation based methodology for evaluating interactive retrieval. This is a very attractive option which has several benefits:

- it is less time consuming and less costly
- it allows control over environmental and situational variables
- it is unaffected by inter searcher inconsistencies
- it allows the comparison of models, and fine tuning can be performed before deploying the real system

Whilst in some senses the process can be less time consuming and costly, this may not necessarily be the case if the software developed to perform the simulation is more costly to produce than running user evaluations. Hence, my proposal is to develop an open source toolkit for developing interactive information retrieval simulations.

This will this ensure that cost involved in performing a simulation is minimized by lowering the overhead to perform simulations. Also, it facilities fair comparisons between different simulations by different researchers. However the real boon is in creating an extensible architecture for interactive information retrieval where the user becomes a distinct entity to model in the process. Modelling the user has been largely ignored in most evaluations, but under a common interactive framework, users (whether they are instantiated as a simulated user, an agent or even a real user) could engage the Information Retrieval System with in different ways. For instance, a flippant web user casually searching for interesting stuff would be modelled differently to a user looking for information about a particular medical condition. Hence, different user models and behaviours can be examined.

Such a project would initially require considerable time to develop an appropriate software architecture that is flexible and extensible enough to cater for the different interactive scenarios that the researchers would like to model. Hence this project would require two to three researchers with interests in developing information retrieval simulations and a software developer to aid in building such a toolkit. The project would take about a year to complete and be extended as required (i.e. instantiating different user models, retrieval functions, types of interaction, etc). The outcome of such a project would be an open source architecture allowing researchers to evaluate their algorithms and interfaces in a simulated interactive environment with different types of user's models.

Using the music in music information retrieval

Mark Baillie

Mark.Baillie@cis.strath.ac.uk

Areas of Interest

Audio-based content classification, video indexing and summarisation, music information retrieval, statistical modelling

Description of Current Research

My PhD work has focused on audio-based indexing of sports video. This is important for efficient video storage, search and retrieval, as well as the development of new interactive video systems. Current annotation processes of video require is expensive, laborious and requires expertise; therefore the automation of this process would be beneficial.

Specifically I have concentrated on two problems; structure segmentation and classification, and event detection. Structure segmentation and classification involves the mapping the video providing an overview of content, similar to the table of content pages found in most textbooks. By doing so an overview of the video content is provided, and also allows the retrieval of specific areas of interest rather than the entire video during querying. Event detection involves locating the key moments in the video. This process is beneficial for summarisation.

To address these problems, I have formally investigated the best methods for parameterising the audio track of video and recognising predefined content using statistical models such as the Hidden Markov model. As part of this investigation I have also addressed such problems as audio segmentation, automatic content analysis and model selection. The final output of the thesis is an indexed sports video that can be efficiently browsed using an interactive video browser.

Project Proposal

A Content-based Music Browsing, Retrieval and Recommendation System
Since the advent of the MP3 audio compression standard there has been an explosion of music downloading across the web. Large online repositories (legal or not) now exist, however current search and browsing of these collections is based on searching meta-data. Often downloaded MP3 files have incomplete meta-data, or typically the searcher cannot remember the artist of song title.
I propose that music can also be searched by content. By building statistical representations for each song, songs can be grouped by similarity allowing for a search by similarity scenario.

- Needed investigation into suitable features that can be extracted for paramterising musical data.
- Investigation into suitable statistical models for acoustic and meta-data.
- Investigation into similarity measures for grouping songs and for querying by content.
- Investigation into recommendation.
- Interactive interface.
- Copyright protection by content analysis

NB. Note that although many systems exist for content-based retrieval of images, little work has been done on the audio portion of the multimedia stream.

Voice Information Retrieval System

Heather Du

heather.du@cis.strath.ac.uk

Areas of Interest

Information access via speech, including spoken query processing, spoken document retrieval, vocal information retrieval, design and implementation of vocal information retrieval systems, user machine interface design for multimedia IR systems

Description of Current Research

My PhD work has been focusing on information access via speech, an area that is associated with the research in information retrieval and speech technology. This research seeks to identify a set of factors that are valuable when users interact with a vocal information retrieval application audibly, in order to seek the desired information objects, and to develop effective ways of exploiting those factors to enhance both the effectiveness and usability when constructing user interfaces for vocal information retrieval systems. I have conducted studies on the differences between queries issued in written form and spoken form in qualitative terms and in terms of their retrieval effectiveness. Written and spoken queries are compared in terms of length, duration, and part of speech. In addition, assuming perfect transcription of the spoken queries, written and spoken queries are compared in terms of their aptitude to describe relevant documents. The retrieval effectiveness of spoken and written queries is compared using different IR models. The results show that using speech to formulate one's information need provides a way to express it more naturally and encourages the formulation of longer queries. Despite that, longer spoken queries do not seem to significantly improve retrieval effectiveness compared with written queries. The qualitative experiment was also carried out in Mandarin Chinese and similar results were found despite this language has completely different semantic structure from English.

Project Proposal

As speech recognition technology continues to grow as a common interface, and existing graphic user interfaces fade away, VIR systems have come into the limelight in today's IR community by offering voice user interfaces (VUI). This implies the design and implementation of systems capable not only of understanding the user's spoken request, finding the required information and presenting it as speech, but also capable of interacting with the user in order to better understand the user information need, whenever this is not clear enough to proceed effectively with the searching. However, while much work has been done in developing dialogue systems which deal with database searching tasks such as flight booking, stock quote and train timetable etc, there is little work carried out on the design of voice user interface for information retrieval systems. The process involved in communicating with a database is far different from that with a document collection.

I am interested in working on the design and implementation of a vocal information retrieval (VIR) system which will enable users to search desired information vocally by interacting with the system. This will involve several technologies to achieve this task. The first ones coming to the mind are speech recognition and speech synthesis. Speech recognition will play an important role for the system to understand the users and speech synthesis will enable the system to communicate back to the users. There will also need a backend IR system that will process the users' information needs. In order to effectively interact with the users, the system also requires a dialogue manager. The dialogue manager interacts in the one hand with user to communicate information events, and on the other hand with the IR component to handle queries and search results.

I am also interested in developing guidelines for dialogue design for information retrieval systems. In traditional searching tasks performed by IR systems, the users are required to issue their information needs by typing into a designated field in a graphic user interface. After performing matching the query terms and document terms based on a specific IR model, the system will present the retrieved documents to users in the form of a summary, or a link, or a combination of both, etc. However, there are lots of open issues when considering a voice user interface for an IR system. For example, queries are not typed but spoken, which means speech recognition component has to cope with a verity of speaking style, speaking rate, accents, and background noises, etc.

The possible outcomes of this project are:

- A study of how to integrate sub components to implement the whole VIR system.
- A study of developing the guidelines for VUI design and tackle issues related to an IR model oriented VIR system.
- A study on the presentation strategies that best convey useful attribute information to users, for example, would users be able to understand the relevance of a document if the content of a document is presented at a faster rate.
- A study on the criteria, users rely upon to choose a document and attributes of the documents that users use as a basis for assessing each relevance criterion.
- Metrics for measuring and evaluating voice user interface.

Academic information management

David Elsweiler

dce@cis.strath.ac.uk

Areas of Interest

Human memory, support tools for human memory, information reuse, interfaces for interactive processing of information.

Description of Current Research

In the first year of my PhD, I have been examining ways in which computer-based tools can be designed to support the human memory systems. My approach has been to gain a better understanding of the human memory, its strengths and limitations and how they affect the way that we work. The aim being the discovery of insights towards the most effective way of providing assistance.

I advocate a user-subjective approach to the management of information (Lansdale 1988). By allowing users to add value to information with their individual perceptions, needs, values and experiences, we can exploit personal connections they have with documents and use these qualities to facilitate the re-finding and reuse of information processed in the past.

I have created a preliminary framework for interaction techniques, exploiting natural human cognitive characteristics that improve memory encoding and facilitate information re-retrieval. In the next two years I hope to formalise and strengthen the framework, in addition to illustrating and evaluating its potential.

Project Proposal

Conducting research requires information of various kinds to be discovered from multiple sources. We learn new facts, opinions and predictions from material read, presentations attended, people we meet and so on. Our familiarity with information and how and where we obtained it, however, diminishes with time, which can generate problems when creating new documents. For example, writing a paper, presentation or literature review may involve pulling together information from sources such as documents, spreadsheets, data analyses, conversations, email messages etc. These creative processes are made even more challenging by disorganisation and poor information management strategies.

I am interested in providing support for the management of literary references and suggest two complementary pieces or work in this area.

Firstly, I would like to conduct a behavioural study of academics and research students that would provide answers to the following questions:

- What different ways do researchers manage the information they process - take notes, keep track of references etc.
- What are the strengths of these strategies?
- When do they fail?
- Do different researchers remember the same type of details about things they have read / information they have processed?
- What items serve as good retrieval cues for the recollection of details of previously processed information?
- People and their work?
- Information content?
- Aspects of search process – how we went about finding particular information?
- Patterns of data exposure, such as trails of exploratory learning (one resource can trigger new avenues of research)

Based on the findings of such an investigation, I would like to develop an application, possibly harmonising with existing systems or methods, to help manage literary references.
Suggestions, ideas for collaboration, or related research are welcome!

Regional factors influencing information access

Magdah Gharieb

Magdah.Gharieb@cis.strath.ac.uk

Areas of Interest

Regional Factors influencing Information Access

Description of Current Research

Information and communication technologies (ICT) are playing an increasingly important role in the everyday life of the society . information seekers have to be equipped with the skills and experience to better take advantage of these new technologies. Many countries have assisted schools, universities, and work stations in providing computers, internet access and electronic mail facilities to support the teaching, learning and working of their population.

Academic libraries are one of the most important dynamic institutions; they aiming at supporting research and education methods. Therefore, they design their web sites to help users get access to resources from anywhere and any time. Substantial developments in the digital field encourage academic libraries to provide different types of information resources to their users. Electronic academic libraries interfaces have facilitated the design of many information access services embracing such as e-journals, databases, e-books, digital libraries and subject gateways. However, these resources can be displayed on different interfaces which create confusion to users who search for user information needs. Accordingly, there are a number of aspects in social environment that could have an impact on getting access to electronic information, such as economic, emotional, cultural, linguistic, and information literacy factors. According to Deschamps, there is a growing gap between developed and developing countries in the easy access to knowledge, information and communications technologies, and using Internet for different activities. This study will investigate the barriers hindering access to electronic information in the digital library age in both different environments.

Objectives and Research Questions

The main objectives of this study will be as follows:

- to identify and classify the differences of the barriers in getting access to electronic information in both Scotland and Saudi academic environments
- to investigate the characteristics of the current user for accessing electronic information and hybrid library services, to assess how the user evaluates the performance of information seeking according to their perspectives and atmosphere.
- Try to find solutions for the barriers to get access to electronic information in both academic systems, that might help academic decision makers to redesign and rearrange their facilities.
- to improve the understanding of information seeking behaviour in relation to electronic information services in a variety of academic disciplines as well as in different academic environments. Also why the different disciplines in the same environment may have the unique problems during accessing to electronic information.
- to improve the understanding of information seeking behaviour in relation to electronic information services in a variety of academic disciplines as well as in different academic environments. Also why the different disciplines in the same environment may have the unique problems during accessing to electronic information.

Aspects of back of the book index/interfaces for digital libraries

Forbes Gibb

forbes.gibb@cis.strath.ac.uk

Areas of Interest

My current interests lie in e-books, digital libraries and language based information retrieval

Description of Current Research

I am working with one PhD student in the area of automated back-of-the-book-indexes (BoBI), with a specific interest in BoBIs for e-books. I am working with another PhD student in the study of meaning in philosophy, information science and information retrieval. I also expect to be working with a new PhD student in the area of multimodal access to digital libraries.

Project Proposal

Future research interests lie in the exploration of some open questions and opportunities regarding BoBIs: what is the optimal size of a BoBI? What metrics should be used for evaluating a BoBI? How useful are meta-BOBIs to collections of e-books; how best should we merge thesauri and BoBIs.

A secondary Areas of Interest is in the development of interfaces to, and the selection of appropriate metaphors for, digital libraries. What, if anything, can we learn from physical library design? How should we present results from digital libraries? What information do readers want from digital libraries? How do they wish information to be organised?

What is it? A metadata service

Srikant Jakilinki

sriks@dcs.gla.ac.uk

Areas of Interest

memories, deep-links, chaos, strands, personal webs, Peer to Peer (P2P) Information Retrieval, metadata accrual, pipelined document-as-query systems.

Description of Current Research

My PhD work has been examining ways in which we can solve the personal IR problem by using digital episodic associative memories (EAM). Specifically, I plan to do this by capturing certain lost interactions between the users and their documents which are aggregated together to extract the 5W's (who, where, when, what, why) for as many document (and their related neighbours) as possible. These additional 'tokens' are added to documents which are then indexed, represented, modelled and queried using EAM principles that we are exploring.

Project Proposal

What I would really like to collaborate on (and parallel to my PhD) is a mechanism to accrue rich metadata for multimedia (like digital photographs) from search engines. The main idea behind is that one can use a web-services or even client/server architecture to extract as much rich metadata as possible and is available. Today, one can send a text query to IMDB or CDDB and get back information about a movie, artists or music file etc. There are online image databases (like, Reuters have heavily tagged pictures) and lots of research search engines which do some neat processing (like giving the number of people or animals or objects in a picture etc.).

If all these search engines or "image-services" can be pipelined and had an open API one could send a document to them, use the document as a query and get the metadata contained in similar documents which could all be tagged/appended to the document. Similarly, other media types could use their own services to get metadata.
Such a seemingly simple project requires considerable changes to the backend. What we are trying to propose is that all researchers should build experimental search engines that conform or make it possible for other clients to query for similar documents. This is no mean task on a global scale but the change at the individual level is potentially small. Synergy is perhaps a good starting point for search engines built in the Glasgow area to be made into collaborative systems.

Adaptable architectures for Peer-2-Peer IR

Iraklis Klampano

iraklis@dcs.gla.ac.uk

Areas of Interest

Peer-to-Peer p2p networking, distributed IR, meta-searching, evaluation methodologies, fusion techniques, p2p simulation.

Description of Current Research

I am working on information retrieval over highly distributed p2p networks. P2p is a newly (re-)invented1 networking paradigm in which the participating nodes are equally capable of providing and using remote services. Researchers expect that IR will be greatly needed in large, highly distributed p2p networks, but the problems involved are many, derived from the open and versatile nature of these systems. The notion of equality enforces that information can be located, potentially, on any participant and that any participant can issue queries. Without the stability provided by the widely-used Client-Server model, all the peers have to cooperate effectively in order a given query to reach these peers that are likely to contain the most relevant content. An additional and very significant problem is the evaluation of p2p IR systems. The reason behind this statement is the difficulty to reflect real application characteristics onto suitable evaluation test beds. My main interests lie within designing appropriate architectures for p2p IR as well as appropriate evaluation settings and strategies.

Position: Architectures for p2p IR. Despite many approaches to the p2p IR problems have been published recently, there are still important problems to be solved in a generally acceptable way. One of the most important ones is the organisation of information providers and the intelligent routing of queries. This problem follows directly from the resource selection problem of the distributed IR field [1]. However, it ends up being significantly different in its p2p IR instantiation since it is typically performed by more than one node (not only the server) and, therefore, a query is expected to be routed into multiple levels of the network topology. Naive content organisation and query routing can lead to inefficient and ineffective solutions in large-scale p2p networks.

Opinions, on tackling the problem of organisation and query routing, divide into database-oriented and content-based solutions. Database-oriented solutions are based on distributed hash tables (for example [4,5], where distributed indexes are built based on hash values of keywords. Alternatively, IR-oriented approaches are based on vector representations of documents and similarity-based routing (for example [2,3]). The first, DB-based, approach can be efficient but it cannot cope with document descriptions of more than a few keywords; otherwise it misuses the network. It also cannot cope with different degrees of relevance; it follows Boolean logic since, in most cases, a query can either result in a hit or a miss.

The second, IR-based, approach can lead to inefficient and non-scalable solutions. This is because informing the network of the various document collections requires high bandwidth, not to mention the maintenance costs due to the unexpected joining and leaving of peers. I consider the problem of locating items of interest (either text documents or other, mainly multimedia, kinds of information), at various degrees of relevance, a challenging and important problem that can be applied in various potential contexts. Therefore, I concentrate my research in content-based solutions.

After our first proposal on p2p architecture[2], we intend to extend our first architecture in various ways. First, the architecture will be re-structured in a modular way. This re-structuring could allow for other solutions to be modelled using a number of base components, therefore providing the research communities with a common base for future collaborations, comparisons of systems etc. Second, we intend to explore thoroughly, through experimentation, a number of widely used p2p IR techniques and models, in terms of their applicability and usefulness in environments. Lastly, and most importantly, we will attempt to improve the information routing capabilities of our system by looking at efficient, effective and incremental ways of describing and communicating shared resources (such as text documents). The outcome of this research will be an intelligent and adaptable architecture for, potentially generic, p2p IR.

[1] J. Callan. Advances in Information Retrieval, Chapter 5 – Distributed Information Retrieval, pages 127-150, Kluwever Academic Publishers, 2000

[2] I. A. Klampanos and J. M. Jose. An architecture for information retrieval over semi-collaborating peer-to-peer networks. In the proceedings of the 2004 ACM Symposium in Applied Computing, pages 1078-1083, 2004

[1]The first version of the Internet followed a p2p fashion.

[3] J. Lu and J.Callan. Content-based retrieval in hybrid peer-to-peer networks. In Proceedings of the Twelfth International Conference on Information and Knowledge Management, pages 199-206, 2003
[4] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A scalable content addressable network. In Proceedings of the 2001 SIGCOMM Conference, 2001 [5] I. Stocia, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnam. Chord: A scalable Peer-To-Peer lookup service for internet application. In Proceedings of the 2001 SIGCOMM Conference, 2001

A real digital library for researchers

Monica Landoni

monica.landoni@cis.strath.ac.uk

Areas of Interest

Human Computer Interaction, Interactive Information Retrieval, evaluation of search engines, visualization techniques, search facilities for e-books and digital libraries

Description of Current Research

I am working on and still involved in a number of major projects including:
E-textbooks for Anywhere Anytime Learning, a study into how to design an e-book reader specifically geared toward education and sub-sequentially evaluate its impact of in virtual learning environment.

Tracking E-book Interactions, a novel approach to gather data about how users interact with e-books.

PENG, an EU Strategic Targeted Research and Innovation (STREP) project, funded under the VI Framework and starting in September 2004, aims at defining a flexible, personalized and context-aware system for the gathering, filtering and presenting news for news professional (i.e. journalists and editors) and general users.

Promoting the Uptake of E-Books in Further and Higher Education, funded under JISC DNER, April to September 2003. Acting as Librarian advisor. This project provided a better understanding of how e-books are being used in higher education and the findings are going to be added and compared to our usability study.

Webkit - Intuitive Physical Interfaces to the WWW (IST 2001-34171). The goal of WebKit is to create an intuitive physical interface to the web, which will enhance the learning process for children. 2002-2004.

Project Proposal

Every year when dissertation time approaches students are asking supervisors or unlucky people with a nice friendly face and an office so small they need to keep the door open in order to breath, THE QUESTION! "Where do I start looking for information about this?" often reworded as "What is the best starting point for my research?" or a number of equally challenging variants.

Personally, I smile to show my understanding, even if the inquiring person has been submitting well documented papers for the past year, has been introduced to the joy of browsing and searching in a real library under the caring eyes of our skilled librarians, and, of course, are more than aware of the delight of Google and its retrieval power. Then, I go on to illustrate how useful information can come in lot of different forms and shapes (even paper!) and how many different sources are available as long as the research question is pondered upon, explored and expanded properly. This is the taste of research!

Nonetheless, there is a clear need for a proper search tool that would allow people to interrogate databases without having to guess from their mysterious names what they are about, as well as the possibility to search papers stored in a library not just by knowing the journal name and issue number or the name of the conference for the proceedings they belong to, but simply by topic or title, really not much to ask is it?

The main aim of the proposed study would be to produce a real digital library for scientists where they could literally find what they are looking for. Without having to fight against rigid data retrieval systems that only return the right answer to those whom ask the right question. The main difference between the proposed system and existing all-the-time-getting-smarter search engines is that our system would be able to search through the wealth of material usually out of reach for normal web crawlers or robot, material academic libraries subscribe to and pay for. Ideally this system would combine quality as assured by the library stamp with high precision and recall plus easy to use and overall standards of usability. It is envisaged that this system could become a challenging arena for a number of researchers involved in different aspects of IR including at least:

- Interactive IR

- Web IR
- Visualisation techniques
- Fusion Techniques
- Metadata and standards
- Communication Protocols
- IR in structured documents
- Summarization
- Filtering techniques
- Personalization

An interesting spin off of this project would be one that offers researchers or research groups the possibility to build their own digital shelf including as well their own publications, with new additions made automatically based on their research profile. Another challenge would be to have a "smart" system that goes beyond what the library offers and mimics what a skilled researcher would do for instance keeping an eye on specific for a or checking well known authors for latest publications. Again all of these systems will be using a search system that allows to really find papers and or relevant projects/initiatives wherever they are.

A starting point would be to look into the catalogue at the Andersonian library and see how searches could become more efficient. Another possibility would be to consider all proceedings published by the TREC initiative and design a system to search across volumes and years possibly using available metadata.

This is indeed a very serious and wide problem this is why I opted for this very light form of presentation to introduce it. Even the definition of the present scenario would per se be a research project, let alone the attempt to find a real working solution. It is a field where our joint efforts are more than necessary, indeed essential. This is an ideal arena for digital library research, possibly working with the Centre for Digital Library Research at Strathclyde University. The e-book research group at Strathclyde will also contribute and possibly complete the study by providing very valuable experience on how to design information to be read and consulted on a screen.

Summarisation across languages

Christina Lioma

xristina@dcs.gla.ac.uk

Areas of Interest

Cross Language Information Retrieval, Natural Language Processing, Computational Linguistics.

Description of Current Research

I am working towards expanding our current information retrieval platform (Terrier) into accommodating cross language information retrieval (CLIR) for a number of Indo-European languages. The morphological analysis of the source and target languages will be assisted by a Part-Of-Speech tagger, as well as restricted naïve query expansion. The specific naïve query expansion method proposed consists in morphologically expanding query terms (after they have been stopped) and matching them against the terms in the documents, as opposed to the mainstream approach which stems document & query terms and then matches their stems.

The proposed morphological expansion will be morph grammatically selective and restrictive, so that the generation of closed class or other 'weak' variations will be disallowed. The retrieval process will be realized by two n-gram models nested within one another. Specifically, a character-based trigram analysis will be nested within a token-based bigram analysis. The character trigram analysis is anticipated to produce interesting results for language sets of common Indo-European origin, mainly on etymological grounds. Lexical bigram analysis has been proposed as a single answer to the problems caused by compounds, collocations, concordances, and multi-word units, which often affect IR performance. The aforementioned investigation will be realized on specific language pairs of common linguistic origin.

Project Proposal

I would like to investigate further the field of monolingual information retrieval for morphologically-rich languages, and particularly non Indo-European languages. Motivation for such an investigation originates from the fact that the said languages are under represented in the fields of information retrieval and computational linguistics, and inversely well represented in the field of theoretical linguistics. Even though these languages have been increasingly making their mark on the Web, the fact remains that today there is a distinct lack of written and spoken corpora available for them, a problem that needs to be addressed immediately.

I am further interested in automatic summarisation technology. Specifically, I believe that there is great potential in combining the knowledge and resources used in cross language information retrieval and multi-document summarisation, in order to address the issue of cross language summarisation. The knowledge to be carried forward from cross language information retrieval, namely language-independent relevance ranking, practically means that unlike conventional automatic summarisation systems, where the focus is in locating and extracting the content-rich parts of the text, post-CLIR summarisation needs to focus on avoiding over-generation and repetition, a task linguistically and computationally simpler than content extraction. Additionally, there is no reason why language variation can impede this type of technology, as the documents to be summarised could be treated as parallel documents, in which case translational equivalences could be easily generated. The sets of documents to be summarised could be selected from the retrieval rankings and clustered in a way that would reflect content shift in the results. Potentially, a cross language search engine could offer the following options to the user:

- rank the results in their original language as per their relevance to the query; or
- cluster content-similar results and generate summaries in the prevailing language, presented as per their relevance to the query.

Designing IR interfaces for young children

Emma Nicol

emma.nicol@cis.strath.ac.uk

Areas of Interest

Human Computer Interaction, Interactive Information Retrieval, Interaction design for children, Search engine evaluation, Digital Inclusion, Electronic Books, Assistive technology and Learner Centred Design.

Description of Current research

My PhD work has focused mainly on investigating the online information seeking behaviour of children. Little work to date has been devoted to children's behaviour with respect to their use of web search engines, with most of it having focused on their use of digital libraries. Several recent studies (Gilutz and Nielsen 2002) (Bilal et al 2001) have shown significant differences in the affective states of children while interacting with Web information from those experienced by adults and have highlighted the poor assumptions on which web design for such children has traditionally been based. With an increasing number of children using the Web for school and homework purposes, my research aims to investigate further how children behave while searching online and what might best be done to assist them in doing so. My research takes a 'learner centred' approach (Soloway et al 1998) and considers not only the searching process but the eventual use of the information that is retrieved. Design techniques such as co-operative enquiry (Druin et al 1998) in which children are involved or 'partnered' at all stages of the process will be explored.

Webkit 2002- 2004 : Intuitive Physical Interfaces to the Web (IST 2001-34171) www.projectwebkit.com. The goal of Webkit was to create innovative tangible interfaces to web information to enhance the learning process for children of school age. My involvement with the project was mainly at the evaluation stage, working with teachers and pupils in mainstream schools and with professionals in various branches of special needs education.

Project Proposal

I have lost count of the number of times I have heard teachers and parents say to children on the look out for information for an essay or other homework task – 'just use Google'. However, almost an equal number of times I have heard them complain that, while the Web is a fascinating place for children to explore and tools such as Google have made it easier to find information, children are often not very good at evaluating the information they find and are even worse at knowing quite what to do with it. These problems are also common to adults of course but in the few studies that have been done (mainly on children's use of digital libraries) these have been shown to be even more pronounced in children. Borgman et al (1995) showed that children have a lack of developed memory recall and several other studies have shown that children are very poor at dealing with cognitive load and are much less focused in carrying out search tasks than are adults (Bilal 2001).

Given these increased cognitive load problems and poor recall memory, both of which have been shown to cause significant searching difficulties in adults, it's a mystery as to why so few search engines have been designed for use by children or have been adapted to include features specifically aimed at helping them overcome these problems. A cursory glance at www.searchenginewatch.com reveals only 3 search engines designed for children with a few of the major web search engines providing special features only in the form of content filtering of inappropriate material - nothing at all which assists with the interactive process.

Studies such as Gilutz and Nielsen's in 2002, showed, for the first time, that much of the web design that has been done for children has been based on assumptions about their likes, dislikes and abilities which are largely false. Children have differing reasons for using the Web from adults and most importantly have entirely different motivations for doing so. There are physical and co-ordination differences to be considered, gender and age differences are more significant, and as well as the cognitive issues already described, children generally are not as technically 'savvy' as many adults assume. Much design continues to be done without taking these factors into consideration, however there are some commercial products on the market that we may be able to learn from.

While working on the Webkit project I was involved in many discussions with experts in the various fields of special needs education, and also with teachers working in mainstream school environment. During these discussions the extent to which assistive software, that is software that has been designed to assist with specific cognitive disorders or difficulties such as dyslexia or dysphasia, became very apparent. There is a vast library of historical educational

software which lies rotting in many school cupboards, unused and unloved, but recently there have been several stories of, apparently at least, overwhelming success.

Software packages such as Clicker (©Cricksoft) and Wordbar (©Cricksoft) to take 2 of the best known examples, which are designed mainly for children with disorders such as dyslexia, employ the use of picture libraries, sound and predictive text to assist children with written activities. These programs are to be found in use in the vast majority of primary and secondary schools in Scotland today and are recognised by many parents, teachers and pupils as providing an excellent means of improving literacy. Surprisingly, despite this success and their widespread use, these programs are largely unknown, unevaluated and unexplored by the HCI and IR communities.

My proposal is to investigate the interaction techniques and styles which these types of software employ to assist children in overcoming their cognition problems, to see whether these can usefully be adapted to inform design decisions made in the design of search engines and search engine interfaces for children (and possibly also adults). The aim being to come up with interfaces which are truly innovative and not, as has been the case in the past, just to create interfaces which are slightly jazzed up versions of the interfaces adults use. ('Just add a few flashy bits and some nice colours and they'll engage with it'). This would involve those with experience in designing search engine interfaces working with professionals with experience of special needs education and in particular those with a background in using assistive software. The Webkit project has forged some good initial links between members of the i-lab group and colleagues in the Faculty of Education at the University of Strathclyde which I feel could be explored further to the benefit of both parties.

References

- Usability of Websites for Children: 70 Design Guidelines Gilutz, Nielsen 2002
- Young Children's Search Strategies and Construction of Search Queries; Druin et al, 2001
- Differences and similarities in information seeking: children and adults as Web Users : Bilal, Kirby 2001
- Children's Search Engines from an Information Search Process Perspective: Elana Broch, Rutgers University 2000
- ARTEMIS : Learner Centred Design of an Information seeking environment for K-12 children Soloway et al 1998
- WEBKIT http://www.projectwebkit.com/
- Cricksoft : http://www.cricksoft.com/uk/

Ideas about multimedia

Reede Ren

reede@dcs.gla.ac.uk

Areas of Interest
Video summarization, video object detection, video segmentation, temporal sequence analysis, visual and audio feature identification and selection, visual and audio information fusion model

Description of Current Research
I have just completed a temporal video segmentation system which divides a football video into a serial of clips based on their video genre, such as game, interview and commercial adverts, and video making skills, i.e. replay and zoom-in. Four classes are identified, namely 'play', 'replay', 'focus' and 'break', which reflect video contents in some sense. Late, these clips are merged into a higher semantic video segment, 'attack' by a hidden Markov model. Another issue is how to employ these video segments in the video browsing and summarization. A trial nonlinear video browser is built for swift scan across game highlights and bringing a summary in respect to user's interaction. Now I am trying to detect game highlights and produce event lights automatically by analysing the labelled video making sequence.

As a key aspect of content based video retrieval(CBVR),video summarization attracts attention from both industry and consumer. It condenses plain videos into exciting briefs while keeping most important issues. Facing the blooming digital video from broadcasting and internet, users depend on the technique to process video data and build up video index instead of current librarian work. There are some obvious issues during the development of an effective video summarization technique, highlight detection, which catches important events or highlights during the video. There are many different methodologies in literature. Chong-wah Ngo(2001) et al classify shots by low level audio and visual features and choose some certain types to take a summary, i.e. motion summary. S.Intille(2001) et al. define a set of heuristic rules to describe highlights based on domain knowledge. Lexing(2001,2002) et al. detect slow motion segments by hidden Markov model and gather them as a brief.

Multi source information fusion model. Video is a compact multimedia, which composes visual, audio and temporal stream and text data sequence. Works in literature can be categorized into two classes, audio-visual based and text information based. Text information based works extract texts from caption, audio and visual stream and then combine them into a text file. It is a simple but reliable information fusion pathway, though ignoring the difference between information source. Audio-visual based ones employ low level features and array audio and visual events by complex models, such as coupled hidden Markov model, hierarchical hidden Markov model and even Bayesian network(MBF). Consider the computing complexity of these models, most reported works follow text information based pathway.

Video summary generation, which organizes highlights and related events to build up a summary. It is something of video browsing, but weighs importance among different highlights and touches the problem how to describe the relationship between highlights and related detailed video segments.

My research focus on the prior two topics, highlight detection and multi-source information fusion. Most video genera, specially sports video, follow certain video making patterns. For example, in football game, video broadcaster employ different zoom-size to catch game details and use 'replay' to show most important moments, i.e. shot. These patterns are something carrying with time and can be described by the mathematical tool, time sequence analysis. For example, Markov chain can treated as the first order of time sequence. I employ these tools to analyse video to detect highlights and decide the data fusion model. It is a new idea on video processing and video summarization. There are a lot of topics calling for hard work.

How people search the web

Ian Ruthven

Ian.Ruthven@cis.strath.ac.uk

Areas of Interest

How people search the web: investigating and categorising user search strategies.

Description of Current Research

Internet search engines such as Google, Lycos and AltaVista provide quick access to billions of web pages. One of the main reasons claimed for the success of systems such as Google is the simplicity of the user interface: to be able to perform a search a user requires no knowledge of how the search system operates. However, the limited support for searching offered by search engines means that the users often struggle to find useful information. Although nearly half the households in the UK have Internet access, Government statistics indicate that one of the major factors in people not exploiting the Internet is that they feel they do not have sufficient skills or confidence to use technology such as search engines. It is a common assumption that experience will give expertise: lots of search practice will develop a searcher's ability to for find information. However this does not accord with a lot people's experience as searchers often have little idea of how to search effectively or how to search differently and many searchers use the same techniques for all searches. As few online searchers have training in information literacy or search techniques it is important to understand what support is required by online searchers, how search experience develops and how aware people are of their own searching behaviour.

To perform a search, users of search engines must make a series of decisions on how to create queries, which documents to view, and how to modify their search requests. Decisions such as these form the basis of information-seeking strategies: approaches taken by individual users to find relevant information. Studies of human information skills have examined how users employ search strategies but these studies have tended to concentrate on one-off investigations or on particular user groups. Thus they do not examine how search skills develop over time and how search skills differ across user populations.

Project Proposal

I am interested in cross-comparing the development of information seeking strategies in different user groups. For example one group could consist of information science students who have explicit training in information literacy, knowledge of search engine technology and 24 hour access to the Internet. Groups such as these often act as experimental subjects in search engine experiments so their information behaviour is important.

Another group could be students with dyslexia. Students with dyslexia and dysphasia often have specific problems with long-term strategies due to difficulties in managing short-term memory. Hence systems that have a high cognitive load may present problems for this group. This group are interesting because they will identify how generic are information seeking strategies: are they as general as other researchers suppose or are information seeking strategies very dependent on the individual?

A third group of may be people who have little or no experience with online search engines, for example people learning to use the internet. This group would provide information on the development of information seeking strategies in groups who have little access, or infrequent access, to online technologies.

There are several possible outcomes of the proposed research.

- A comparative study of information seeking strategies and skills across three diverse user groups.
- A study of how information seeking skills develop over time
- Metrics for measuring information seeking skills in terms of aspects such as the quality of skills, or the flexibility of skills.
- Information on what kind of interface support do users require to employ strategies
- A study of how information-seeking skills affect experimental studies of interactive search engines.
- It is common to record demographic information about experimental subjects but most people simply record this information they do not analyse its impact on the experimental results
- Understanding requirements for supporting the development of good information seeking skills and the strategies that successful users employ

Framework for context aware IR

Simon Sweeney

simon.sweeney@cis.strath.ac.uk

Areas of Interest

Mobile information retrieval, results presentation, personalisation, summarisation, user studies, use of contextual information.

Description of Current Research

My PhD work is centred around supporting information access on mobile devices, in particular, focusing on presentation of search results for such devices. Results presentation for mobile devices (phones, PDAs, laptops) should be personalised and context dependent. There are several parameters to personalisation, particularly the case when the user is mobile. As mobile devices are by definition personal devices one of the dimensions of personalisation is the device itself with it's associated characteristics (limited bandwidth, input facilities, screen size). Typically content for small screen devices has to undergo some form of processing for optimal viewing on such devices [0, 0]. One means of adapting search results presentation is to employ summarisation techniques, the aim being to summarise the results with minimal loss of user perception of relevance. Indeed, some forms of summarisation can improve user perception of relevance, using for example query-biased techniques [0]. We have carried out some experiments investigating the effects of results presentation and screen size (phone and PDA) using query-biased summaries [0, 0]. Results indicated user preference for, and better performance with shorter concise summaries that were relatively brief, 7% of the document length (up to a maximum of 3 sentences). I am currently investigating if there is a way of finding an optimal a-priori summary size, given the device screen size. This would address the "cold start" problem, since personalisation would enable a user to change summary size to suit their preference after an initial viewing. Further dimensions that I intent to explore includes approaches for personalised summarisation. There are many ways to produce summaries and an effective way, for personalisation, would be to learn from the user how to make the best summary for their particular needs, or to suit their interests. This is the stage where I am currently focusing most of my efforts.

The overall objective for my PhD is to develop a system that will allow me test a number of prototypes for different platforms (Mobile phone, PDA, Pocket PC, laptop) in a user-orientated task-based environment (as a user study). This will allow me to make an assessment of the automatic summarisation technologies employed and to indict any limitations for their use in this type of application. Results from the study will also hopefully provide some feedback on possible improvements/extensions that can then be implemented and re-evaluated.

Project Proposal

Context-aware retrieval (CAR) can be viewed as IR that is engaged and driven by sensors responding to a user's current activity. Whilst the user is engaged in an activity and resides in, or is moving around some environment information is retrieved, and is available for viewing, that is deemed relevant to the user's current context [0]. As a framework CAR seems an ideally suited for supporting information access to mobile devices. An extension of my current work that focuses on the presentation of results may instead concentrate on the retrieval and delivery aspects of supporting information access to mobile devices using CAR as a framework. Where some interesting research questions may include, how to integrate personalisation information into the retrieval process? What effect does this have on retrieval performance?

Document model as a form of summarisation: In terms of investigating new summarisation approaches that can be combined with personalisation then a query biased language model approach may prove useful. After an initial meeting with Leif Azzopardi following SIGIR'04, we have discussed the possibility of a collaborative effort to develop a user study experiment to evaluate and compare a language model approach to other summarisation approaches. The outcome of which may provide some interesting results.

User study design

A further area that collaboration may be possible is in the area of user study design. This is an important part of my PhD work since I am interested in issues associated with the design of user-orientated task-based user evaluations/studies.

References

P. Brown and G. Jones. Context-Aware Retrieval: Exploring a new environment for information retrieval and information filtering. Personal Ubiquitous Computing 5(4): 253-263, 2001.

M. Jones, G. Marsden, N. Mohd-Nasir, and K. Boone. Improving Web Interaction on Small Displays. In Proceedings of 8th WWW Conference, Toronto, Canada, May 1999.

M. Jones, G. Buchanan, and H. Thimbleby. Sorting Out Searching on Small Screen Devices. In Proceedings of the 4th International Symposium on Mobile HCI, pp. 81-94. Springer, 2002.

S. Sweeney, F. Crestani, and A. Tombros. Mobile Delivery of News using Hierarchically Query-Biased Summaries. In Proceedings of ACM SAC 2002, pp. 634-639, Madrid, Spain, March 2002.

S. Sweeney, and F. Crestani. Supporting Searching on Small Screen Devices Using Summarisation, Mobile HCI 2003 International Workshop, Mobile and Ubiquitous Information Access, Udine, Italy, pp. 187-201, September 8, 2003.

A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In Proceedings of ACM SIGIR, pp. 2-10, Melbourne, Australia, August 1998.

Position Statement

Jana Urban

jana@dcs.gla.ac.uk

Areas of Interest

Multimedia retrieval, Content Based Information Retrieval (CBIR), interaction and interfaces, relevance feedback, user studies.

Description of Current Research

While a lot of attention in CBIR has been drawn to the underlying techniques, such as feature extraction, indexing, matching and not to forget relevance feedback, there is still much scope for improving the interaction paradigm between the user and the system. A system that tackles the intrinsic problems of CBIR, such as the semantic gap, the query formulation problem, and long-term or time-varying information needs, should place searching in a personal, task-related context. I'm developing a "search in context system", EGO (Effective Group Organisation), which deals with these problems by combining the search and organisation of image collections.

In EGO, the user engages in an interactive organisation process, which may span several sessions. The interface assists the user in grouping images by providing recommendations. As a consequence, a semantic representation that reflects the user’s mental model of the work task will emerge. The system learns from the organisation adapting to the user's requirements. Thus, EGO provides an environment for the user to organise and locate images for their day-to-day requirements.

Implicit methods for searching interests

Ryen White

ryen@umiacs.umd.edu

Areas of Interest

User interaction and modelling, search result presentation and visualisation, adaptive search systems, Web document summarisation, experimental design and analysis, affective computing

Description of Current Research

Searchers can find the construction of query statements for submission to Information Retrieval (IR) systems a problematic activity. These problems are confounded by uncertainty about the information they are searching for, or an unfamiliarity with the retrieval system being used or collection being searched. On the World Wide Web these problems are potentially more acute as searchers receive little or no training in how to search effectively. Relevance feedback (RF) techniques are allow searchers to directly communicate what information is relevant and helps them construct improved query statements. However, the techniques require explicit assessments that intrude on searchers' primary lines of activity and as such, searchers may be unwilling to provide this feedback. Implicit feedback systems are unobtrusive and make inferences of what is relevant based on searcher interaction. They gather information to better represent searcher needs whilst minimising the burden of explicitly reformulating queries or directly providing relevance information.

The title of my Ph.D. thesis is 'Implicit Feedback for Interactive Information Retrieval' and the techniques I have proposed aim to increase the quality and quantity of searcher interaction, using this interaction to infer searcher interests. I develop search interfaces that use representations of the top-ranked retrieved documents such as sentences and summaries to encourage a deeper examination of search results and drive the information seeking process.

In my thesis I present implicit feedback frameworks based on heuristic and probabilistic approaches. These frameworks use interaction to identify needs and estimate changes in these needs during a search. The evidence gathered is used to modify search queries and make new search decisions such as re-searching the document collection or restructuring already retrieved information. The term selection models from the frameworks and elsewhere are evaluated using a simulation-based evaluation methodology that allows small differences in model performance to be detected.

The thesis describes a number of user evaluations which test the interface components I create and the implicit feedback frameworks with human subjects. The largest experiment involved 48 subjects with different skill levels and search experience. Results from this experiment show that searchers are happy to delegate responsibility to RF systems for relevance assessment (through implicit feedback), but not more severe search decisions such as formulating queries or selecting retrieval strategies. Systems that help searchers make these decisions are preferred to those that act directly for them or await searcher action.

More information on my research, including my publications, can be found at http://www.dcs.gla.ac.uk/~whiter/.

Proposal

I will soon be leaving to Glasgow to begin a position as a postdoctoral research associate in the University of Maryland Institute for Advanced Computer Studies, Maryland, USA, working with Professor Douglas W. Oard. Initially I will be working on two research projects: (i) MALACH (Multilingual Access to Large Spoken Archives) (ii) an as yet unnamed project on techniques to search conversational text (e.g., electronic mails and USENET postings). I will be collaborating with researchers in the Cross-Language Information Retrieval and Human-Computer Interaction Laboratories (the latter established and partially run by Ben Shneiderman) and the Department of Computer Science. I also hope to find time to further my own research on implicit methods for inferring searcher interests and other aspects of interactive information retrieval.

Grid Computing

Fabio Simeoni

Fabio.simeoni@cis.strath.ac.uk

Areas of Interest

IR on the GRID, DIR, Digital Library, Architectures, Metadata Interoperability

Description of Current Research

Framed in the context of diligent – a large scale research project of the European community focused on the integration of digital library technologies and grid technologies. My main research focus is on the migration of current approaches to distributed information retrieval to the large scale and performance orientated assumptions underlying grid platforms. The stress is thus ion the architectural issues for IR and on the injection in the field of service orientated computing principles.

Scale Free Networks in Web Retrieval

Claudia Hauff

claudiahauff@gmail.com

Areas of Interest

Web JR, link structure analysis, scale free networks.

Description of Current Research

I am looking at how to exploit the knowledge that many real networks have a power law degree distribution. In the specific case of the web, the SFN model allows us to determine the number of incoming hyperlinks a page should have, that is used to predict the popularity of it. At the moment this approach has only been tested on a web collection, I hope to apply this approach to other networks as well in the future.

Position Statement

Iadh Ounis

Ounis@dcs.gla.ac.uk

Areas of Interest

Web IR, Web search engines, Implementation of large scale IR systems, distributed IR and Intranet Search, Probabilistic models for IR, Cross Language IR, semantic web and IR.

Description of Current Research

Terrier: A platform for building large scale IR applications.

Probabilistic Models for IR: Models that are derived from empirical data and adapt to the users information needs and queries. These models are based on the Divergence from Randomness Framework (DFR).

Web Search: Anything related to web retrieval and evaluations. From combinations of evidence to systems architectures and from user modelling to evaluation.

Multi-Lingual Distributed Intelligent Tutoring System

Abhishek Sharma

abhishek@dcs.gla.ac.uk

Areas of Interest

Cross Lingual Information Retrieval, Natural Language Processing, MDITs.

Description of Current Research

Currently working on developing a interface obviously supporting cross lingual queries. Basically for the terrier which is the search engines our university website uses. So would like to use Babylon Dictionary (glossary files actually) and would like to initiate my work with English and French queries.

Proposed Work

Would like to develop MDITS; Multi Lingual Distributed Intelligent Tutoring System which will be using the research from Cross Lingual Information Retrieval.

Position Statement

Di Cai

caid@dcs.gla.ac.uk

Areas of Interest

Discriminative Information, Query information, thesaurus normalisation, semantic relations.

Description of Current Research

- To measure the power of discrimination of terms based on Information theory.
- To improve precision performance by query reformulation.
- To normalise thesaurus for analysing the semantic relations between based on set theory and evidential theory.
- To measure the mutual information from analysing statistical relations between terms.

Web Information Retrieval

Vassiliis Plachouras

Vassiliis@dcs.gla.ac.uk

Areas of Interest

Web information retrieval, large scale test collection experiments, selective application of retrieval approaches.

Description of Current Research

My research is currently focused on web information retrieval; more specifically I investigate the selective application of different retrieval approached per query based on evidence from the set of retrieval documents.

The Application Process

Norma McNaught and Deirdre Kelliher

n.mcnaught@enterprise.gla.ac.uk, d.kelliher@enterprise.gla.ac.uk

The content and quality of the application you submit will determine whether or not you are successful. Therefore it is vital that you have a full understanding of what is required, as well as knowing the various stages of the application process, so that you maximise your chances of gaining an award.

Careful attention will help you to avoid some of the basic pitfalls and improve the funding chances of your research idea.

1. Allow yourself time. Preparing a draft proposal and consulting on it, preparing the project costings and getting advice on these, eliciting the necessary approvals and signatures from the institution at the end of the process as well as reading the regulations of the grants scheme to learn what is and what is not permissible, are all time-consuming parts of the process of application.

2. Study your funding source - all funding agencies will have their own criteria for deciding on allocation of their resources. It is worthwhile taking time to familiarise yourself with these and ensuring that your application clearly addresses your targeted source of support. The 6 UK Research Councils are funded by the government with an overall mission "to promote and support by any means, high quality, basic, strategic and applied research and related postgraduate training; to advance knowledge and provide trained researchers who meet the needs of users and beneficiaries, thereby contributing to the economic competitiveness of the UK, the effectiveness of public services and policy, and the quality of life; and, to provide advice on, and disseminate knowledge and promote public understanding of, sciences". Four characteristics of all successful research grants are constant. They must:

- promise excellent research
- be of value to potential users outside or within the research community
- convince of the ability to deliver research
- demonstrate value for money (not necessarily the same as cheapness).

3. Read the rules and the guidance notes attached to the application form which are designed to help you through the 'filling in' process. This cannot be over-stressed; familiarising yourself with the content of the ESRC funder's Guidelines may seem tedious but will help you to avoid basic mistakes which at best will require clarification with office staff and at worst may prejudice chances of success. It is also a good idea to take photocopies of the application form to do the drafting of the more detailed sections. Make sure you are using the current versions of the application form and Research Funding Guidelines. If in doubt check with the office staff at the Council.

4. Discuss your application with peer groups, colleagues and, if you are a relatively new researcher, with senior and more experienced researchers. Experienced collaboration or supervision rarely goes amiss. If you have never sent in an application before try to get the advice of someone who has already been successful. Contact the people you intend to nominate as referees and make sure they know what you are doing. It is not uncommon for nominated referees to be unaware of the substance of the work they are asked to comment on, have little knowledge of the applicant or his/her work, or give a very poor grading. Some have even been known to decline to comment!

5. Justify your costings which should be considered with care and close reference to the ESRC the funder's guidelines. Be realistic - lavish costings are unlikely to find favour with the Board and a proposal which promises the earth at remarkably low expense will be regarded with caution. Applicants should think carefully about the time and resources needed to complete the research successfully within the specified period. Awards will be based on the eligible costings included in applications (unless otherwise agreed by the funder) and will be subject to standard indexation and cash limited at the time of announcement (UK Research councils) so it is important to get costings right when applying. A well thought out financial plan helps to create confidence in the proposal generally. Give as detailed a breakdown of costs as possible so that the Board can properly assess the case for support. Do make sure that what you are asking for is allowed within the regulations. Bear in mind that the funder is looking for value for money.

6. Content and Presentation The research proposal is the means by which you will be trying to convince the Board that your application is worth funding so think carefully about what information you are going to give and how it is presented. Make sure you think your plan through and cover all stages.

The Application Process - Questionaire

Ask yourself the following questions.

- Have I clearly formulated the problem, have I put it in context of contemporary scientific and theoretical debates, demonstrated the way in which my work will build on existing research and make a contribution to the area? Is there a clear and convincingly argued analytical framework?
- What will the research do, to whom or to what, and why?
- Have I established appropriate aims and objectives? Are they clear and concise, do they reflect intellectual aims and practical, attainable objectives?
- Have I provided a well-thought out research design in which there is a reasoned explanation of the scale, timing and resources necessary? Am I being realistic about these? Am I using the most relevant approach and the most appropriate methods? How will it relate to and deliver the objectives?
- What will my research design allow me to say in the interpretation of anticipated results?
- Have I given a full and detailed description of the proposed research methods? Is there any innovation in the methodology I am planning to use? Am I developing any new methods or using established methods innovatively?
- If I am using data collection have I considered already existing data resources? Am I sure that access will be given where necessary, and do I have written confirmation of this? Am I convinced of its quality, validity, reliability and relevance? Have I considered the costs of cataloguing and preparing data for archiving?
- Have I demonstrated a clear and systematic approach to the analysis of data and how this fits into the research design?
- Have I thought about the ethics of what I am planning to do? Are there any sensitive issues or potential problems which need to be addressed? Have I fully consulted on these issues and obtained the approval of an ethical committee where required.
- Have I recognised and planned for the skills and competencies that will be required to bring the work to a satisfactory conclusion?
- Have I anticipated potential difficulties? Have I shown that I recognise these and discussed how they would be handled.
- Have I provided a bibliography? This will be used in the selection of referees and will indicate your familiarity with the theoretical grounding and current state of the art of your subject. Where there is genuinely little or no relevant literature, explain this fully. Funders and referees will not assume your erudition, they want evidence.
- This proposal will be subject to the critical appraisal of my peers. Am I satisfied that I have fully defended my chosen research design and made it clear why others are not appropriate?
- Have I identified potential users of this research outside of the academic community; have I involved/consulted them in my planning? Have I arranged for their continuing involvement in the research process in an appropriate way?
- Have I considered the possibility of co-funding of the research, with this funder being asked to provide only a proportion of the project funding?
- The Application Process
- Have I provided a clear dissemination strategy for the research demonstrating how the research outcomes will be communicated to all interested parties including potential users of the research outside of the academic community?

Convey to the funder your genuine interest, understanding and enthusiasm for the work. Keep the following questions in mind as you plan:

- what is the story you are telling,
- what is the audience,
- why does it matter,
- why now,
- why you!

Six pages are (normally, for Research Councils) allowed for the research plan so do not provide the equivalent of twelve by filling the space with tightly packed typescript. This looks unapproachable, is difficult to read and suggests a lack of clarity in the mind of the writer; proposals using a small type face (smaller than that produced by point 12) will not be accepted. Conversely, a proposal consisting of half a dozen well spaced paragraphs covering only 2 or 3 pages, usually confirms the suspicion that not much is on offer. It is also important to make sure that you devote enough space in the proposal to describing the research you intend to conduct and the research design and methods - the Board finds it very frustrating when applicants devote pages to explaining why their proposed research is exciting but then provide only a short and inadequate explanation of how they propose to explore this in practice.

Write in plain English. Your proposal is likely to be seen by a great many people, some of whom will not be versed in your particular specialisation. Detail and specification may necessitate the use of disciplinary or technical terminology and this will be clear to peer reviewers, but the ideas you wish to convey and your reasons for doing so should be apparent to a wide audience. By the same token, do take the trouble to check spelling, grammar and punctuation. These are all part of the quality of presentation and presentation matters!

7. Dissemination Funders will usually place emphasis on ensuring that researchers engage as fully as possible with the users of research outcomes. These may be other academics, government departments, public bodies, businesses, voluntary organisations or other interested parties. Try to consult with and involve people who could make a valuable contribution to the research and who could provide support and interest. Try to do this in the planning of the project and build dissemination activities into the structure of your research plan rather than give them passing reference as an after thought at the end.

8. Check the details - once you have completed the application form make sure that all the required information is provided. Some of the most common omissions and problem areas are:

- obtaining all the necessary signatures and institutional stamp (not required if submitting using Electronic forms which must be despatched by registered despatchers in institutions – i.e. Grants Managers at GU),
- a covering letter in the case of resubmissions,
- omission of dates of birth for co-applicants or of cvs for named research staff
- the correct number of copies (not required for applications submitted using electronic forms),
- a realistic start date,
- The Application Process
- details of previous/current applications with reports on current projects or end-of-award reports where required. Funders will generally not process new applications if an end-of-award report is overdue,
- a proposal limited to the number of pages defined by the funder.

9. If you are successful after all the hard work, planning and nail-biting, then congratulations, and we hope the work proceeds without too many problems. However, if difficulties arise such as delays in recruitment, staff illness, replacements, or changes to the work plan then please let the funder know immediately.

10. If you are unsuccessful your application will fall into one of two categories:

a) a proposal graded alpha but not funded. This means that although your proposal was one which the Council would have wished to support in principle, there were insufficient funds available for it to do so. This is bound to be disappointing but unfortunately the Councils are only able to fund approximately a third of alpha-rated proposals. This is stiff competition by anyone's standards! Even if you have received an alpha grade do not assume that this means you can resubmit the same proposal with some window-dressing adjustments. In view of the intense competition and the large number of new alpha-rated applications which Councils cannot fund, some Councils will not normally accept resubmissions of unsuccessful applications. Exceptionally, resubmissions may be considered where it can be demonstrated that the proposal has been substantially revised and the changes made summarised in a covering letter accompanying the application;

b) a proposal graded beta or reject. If you did not get an alpha grade then the referee and assessor comments may offer some helpful guidance but you really need to think carefully about the quality and value of the work you have proposed.

11. Finally - we hope you have found these notes useful and wish you success with your application.