ACM SIGIR 2002 – August 11th to 15th 2002, Tampere, Finland

Review of SIGIR 2002

Author: A. MacFarlane, City University, London

Introduction

The 25th Annual conference on research and development was held in Tampere Hall, Tampere, Finland. In all, 2 keynote talks were given, 44 papers presented, 41 posters delivered and 12 demonstrations of various software were given. A vast array of subjects in information retrieval was presented from high-level user studies to theoretical aspects of IR to implementation of IR systems. We review each of these areas below. We give a summary of the talks for each session, putting those sessions that are closely linked together. For each paper presented we provide the main author, the title of the paper and the pages where full paper can be found in the proceedings.

Keynote Talks

Prof K. Van Rijsbergen of *Glasgow University* gave a personal perspective on the history of information retrieval entitled "Landmarks in information retrieval: the message out of the bottle" (p1). He posed the question: why has the subject survived? It was felt that the reason for this was that a strong experimental methodology had been developed over the years through the Cranfield, TREC and Medlars experiments to name but a few. The history of IR from its early days was also given e.g. Fairthorne, Luhn, Mooers etc: 1958 being the approximate starting point for the discipline as we understand it now. Various important landmarks that have importance today were outlined including inverted files to improve efficiency; the vector space and probabilistic models with their various weighting schemes to improve effectiveness: Relevance feedback to give use control over output.

Prof Kimmo Koskenniemi of the *University of Helsinki* gave a talk entitled "Is natural language an inconvenience or an opportunity for IR?" (p315). The motivation for this talk was to build links between the computational linguistics and information retrieval communities. He stated that thinking and language are not easily separated and that language contains ambiguity and vagueness which can often hamper effective communication between people. A fair amount of time was spent discussing aspects of language such the complexity of some of them; what is a language and how do they change etc. A further problem: grammar does not encode full meaning. It was stated that languages change in order to make them 'better', to adapt to changes in the environment. The basic question set by the talk was how do we use computational linguistics in information retrieval to solve problems of complexity, lack of information and ambiguity?

Web Information Retrieval

The session on Web IR covered various areas including data fusion techniques and efficiency. The paper delivered by *Anh & Moffat* entitled "Impact Transformation: Effective and Efficient Web retrieval" (p3-10) discussed the use of pruning techniques in term weighting schemes to increase both retrieval effectiveness and efficiency for short queries. Their average precision results demonstrate that the method described is superior to web track participants [2]. However the method has a detrimental impact on longer queries. *Park et al* in their paper entitled "Analysis of lexical signatures for finding lost or related documents" (p11-18) focused on the problem of broken links and on how to improve the persistence of information on the web. They proposed a method called "lexical signatures" which is a short list of identifying words from a given document. It is asserted that a robust hyperlink can be created using such methods. Various methods of lexical signatures are examined and it was demonstrated that each method has its own advantages/disadvantages: no one method of lexical signatures was best in all situations. *Si and Callan* discussed the merging problem in distributed IR in their paper entitled "Using sampled data and regression to merge search engine results" (p19-26). The method uses regression techniques to estimate the centralised scores (much less expensive than centralised search itself). This is done by collapsing sample documents into a 'centralised' database which allows estimation. Results show an increase over the CORI method both in terms of retrieval efficiency and effectiveness. Finally *Kraaij, Westerveld and Hiemstra* discussed the issued of entry page search in their paper entitled "The importance of prior probabilities for entry page search" (p27-34). There are various sources of evidence that can be use to find a home page e.g. page length, incoming/outgoing links, URL form (root, subroot, path, file). It is asserted that entry page search is different from ad-hoc search, and any system that uses straight ad-hoc techniques performs disappointingly. Their conclusion was that the URL form method proved to be strongest single component, but that it needs to be combined with other features to obtain the best results.

Information Retrieval Theory

The IR theory session concentrated on the use of language model techniques. *Hiemstra* in his talk entitled "Term-specific smoothing for the language modelling approach to information retrieval: the importance of a query term" (p35-41) described some problems in current IR theory which are not dealt with very well such as stop words, coordination level ranking etc. For example IT (i.e. case-folded to 'it') will be a stop word in many systems, but may refer to the Information Technology concept in a particular document. Various techniques for smoothing probabilities are described in order to emphasis the importance of a term that may address the problems outlined. *Rong Jin et al* described a new language model in their paper "Title language model for information retrieval" (p42-48) called 'title language models'. This model is based on the following concept: queries are similar to titles, short in length and concise in description. A title language model is estimated from a document, then the likelihood that the query

would have been generated from this estimated model. This can result in a sparse data problem that can be handled by smoothing (as in *Hiemstra* above). Results using TREC data show that the title language model outperforms traditional language models and the vector space model significantly. Lastly *Zhai and Lafferty* described "Two-stage language models for information retrieval" (p49-56). The problem they outline is the setting of retrieval parameters according to user preferences. Can we adapt parameters automatically, thus allowing some degree of flexibility for users? The two stages of their smoothing model capture different aspects of the query, providing optimal setting of values for constants. Their results show that retrieval effectiveness is a least equal to and can be superior to single stage methods.

User Studies

The session on user studies contained three contrasting papers the first of which (by *White et al*) described an evaluation of two schemes for finding relevant documents using top ranking sentences (p57-p64). The two techniques studied are; presentation of sentences ranked highly by the user's query and evidence captured from user interaction in order to adapt the interface. Evidence shows that both methods have an effect on retrieval effectiveness. *Chen et al* in their paper "Predicting category accesses for a user in a structured information space" (p65-72) pose the question: how do you capture shifts in the interests of users? They use a two-phase model, the first of which generates a list of user preferences and a second which makes a prediction on these user preferences. Results show that analysing the whole history of user access yields the best results. *Smith* delivered a paper entitled "Detecting and browsing events in unstructured text" (p73-80) describing the use of geographical and temporal information in browsing to obtain evidence on the significance of various events e.g. interests in certain areas of the world at a particular time identifying of areas likely to be of interest to the user.

Filtering and Text Categorisation

There were four sessions in this area of interest namely: *filtering*, *collaborative filtering*, *text categorisation* and *clustering*. The filtering session started with at talk entitled "Novelty and redundancy detection in adaptive filtering" (p81-88) by *Yi Zhang et al*. The user wants relevant and *novel* information as soon as it arrives. Novel meaning a document is not similar to those previously delivered. The focus of their research is therefore on redundancy, and on using both redundancy and relevance detection to improve effectiveness. They show a number of successful methods for redundancy detection. *Leuski & Allen* in their paper "Improving realism for topic tracking evaluation" (p89-96) focused on the tracking element of the TDT task, in a real interactive session. How many documents are users prepared to read so that the system can use relevant documents to improve effectiveness? There will be a limit on the number of documents users are prepared to read. However, they demonstrate that tracking can be done interactively. *Chai et al* in "Bayesian online classifiers for text classification and filtering" (p97-104) discussed the use of Bayesian methods to classify documents. They show that their implemented methods (namely Perceptron and Gaussian) are comparable to the best methods for text classification, namely Support Vector Machines.

The text categorisation session consisted of four papers which started with a paper entitled "Unsupervised document classification using sequential information maximization" (p129-136) by *Slonim et al*. They use a simple sequential clustering algorithm with a linear time complexity and where the number of clusters governs the space complexity, not the number of documents. Their method is superior to all other clustering methods on small to medium sized collections, and comparable to a *naïve bayes* method. *Kawatani* in "Topic difference factor extraction between two document sets and its application to text categorization" (p137-144) looked for distinctive information in a class in order to improve the process of text categorisation. Documents are represented as vectors, and the method put forward finds axes in these vectors that distinguish documents. An improvement was demonstrated on a small collection. In "Text genre classification with genre-revealing and subject-revealing features" (p145-150) *Lee & Myaeng* described methods for genre based classification using information such as stylistic characteristics, reportage, fiction, review, legal document etc. Results can be enhanced using this method by treating genre as a feature. *Crammer & Singer* in "A new family of online algorithms for category ranking" (p151-158) described methods of ranking categories with respect to documents. The demonstrate improvements in performance on both small and large Reuters collections over Rocchio and Perceptron algorithms.

The clustering session contained three papers starting with *Liu et al* in a talk entitled "Document clustering with cluster refinement and model selection capabilities" (p191-198). Their aims are to achieve accuracy of clustering and the ability to estimate the number of clusters in the given corpus. They use an iterative method that examines the feature set in order to obtain an optimal set of cluster by convergence. Improvements in performance are shown. *Pantel & Lin* in "Document clustering with committees" (p199-206) discussed the CBC (Clustering by Committee) method, which increases the intra-group similarity while decreasing the intra-group similarity. They also describe a new evaluation technique for measuring cluster quality by examining the difference between clusters output by an algorithm and manually constructed classes. In "Probabilistic Combination of Text classifiers using reliability indicators: models and results" (p207-214), *Bennett et al* put forward a probabilistic method which combines the output of several clustering methods – called metaclassifiers – in order to improve the quality of output clusters. Reliability indicators (variables) are used show why different classifiers are better for disparate situations.

Last of these session was on collaborative filtering which contained three papers on differing aspects of the area. *Canny* in "Collaborative filtering with privacy via factor analysis" (p238-245) addressed the problem of retaining privacy in various situations such as peer to peer and peer to server. It is asserted that anonimising logs is insufficient to protect the identity of users, as evidence from logs can be used to pinpoint users by their activity. The use of encrypted communication of user activity and factor analysis can be used to protect the privacy of users. The issue of implementation in collaborative filtering is tackled by *Coster & Svensson* in "Inverted file search algorithms for collaborative filtering" (p246-252). They showed that inverted files are faster than the in-memory vector method by a significant amount. "Methods and metrics for cold-start recommendations" (p253-260) were discussed by *Schein et al*, who concentrate on the use of augmented aspect models to cold start collaborative filters. Cold start is the

situation where no user recommendations have been made and new items have to be chosen in order to initialise the filter. They have a new method for evaluation which they claim is general and can be applied to situations in collaborative filtering other than cold start.

Summarization

This session contained three papers with three distinct themes, but which use extraction based methods. *Amini and Gallinari* in their paper "The use of unlabelled data to improve supervised learning for text summarisation" (p105-112), described the use of unlabelled training data on a supervised machine learning algorithm to improve the performance of summarization. They were able to show distinct improvement using a larger collection, but were unable to do the same for a smaller collection. *Zha* in their paper "Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering" (p113-120) described the use of clustering algorithms together with a mutual reinforcement principle for simultaneously extracting keyphrases and general text. *Hardy et al* in "Cross-document summarization by concept classification" (p121-128) described experiments in the Document Understanding Conference (DUC) using a system called XDox. XDox has a two-stage summarization process which uses passage clustering. Examples of output are given in the paper which are clearly readable.

Cross Language and Arabic IR

The Cross Language or CLIR session contained four papers, three of which dealt with the problem of query translation, while the other focused on relevance models. The emphasis of *McNamee & Mayfield's* work in "Comparing cross-language query expansion techniques by degrading translation resources" (p159-166) was on the evaluation of query expansion techniques in order to rectify errors in translation across sources of variable quality. They are able to show modest performance gains, while putting forward an explanation as to why there is conflicting evidence in the literature. In "Statistical cross-language information retrieval using N-Best query translations" (p167-174), *Federico & Bertoldi* used two statistical models, one a query translation model and the other a query document model. The results of these two models are fused choosing the most probable translation. The issue of translation selection was discussed by *Gao et al* in "Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations" (p183-190), who show that translating triples is more effective than using word-to-word methods. *Lavrenko et al* proposed a method in "Cross-lingual relevance models" (p175-182) which utilises language model techniques that do not rely on translation of either the document or query. They use disambiguation and query expansion techniques in a topic model that can be estimated from a parallel corpus or a dictionary. The model achieves a performance that is 95% of what was recorded on monolingual runs.

In the Arabic IR session three papers were presented on subjects as diverse as term selection and stemming: the focus was on problems in the language which have to be tackled by IR systems. The problem of OCR was addressed in "Term selection for

searching printed Arabic" (p261-268) by *Darwish & Oard*. Many texts in Arabic are available only in print. They discuss mechanisms for reading in such texts via a scanner and handling diacritics, stems etc together with various schemes such as n-grams. The best combination of strategies was 3-grams based on characters, together with light stemming. They have built a test collection of 2.7k scanned documents, together with 25 hand built queries with exhaustive relevance judgements. *Xu et al* in "Empirical studies in strategies for Arabic retrieval" (p269-274) outlined some challenges which occur due to the highly inflective nature of Arabic: words may have multiple stems; variant spellings – diacritics; the prevalence of synonyms; significant problems with word ambiguity to a tri-literal root system. In order to tackle these problems they put forward a mechanism that is table driven and uses a list of valid prefixes, stems and suffixes. They also describe an English/Arabic parallel corpus that is used to develop thesauri: the assumption made is that synonyms in one language will be translated to much the same words in another language. They were able to boost performance by about 18% using their thesauri method in one particular experiment. In "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis" (p275-282) *Larkey et al* concluded that stemming for Arabic is very important. They take a number of stemmers and morphological analysers and compare their effectiveness on the TREC Arabic collection. The best results were achieved with increase in weak stemmer strength and morphological analysis.

Queries

The four papers in this session looked at a variety of issues in queries such as query refinement, question/answering, the prediction of query performance and reducing query ambiguity. *Carmel et al* in "Automatic query refinement using lexical affinities with maximal information gain" (p283-290) described methods of refining a users query by automatically adding terms related to original query terms: these are termed lexical affinities (LA's). The terms may be closely related to each other by being adjacent in the corpus for example. Results show that adding LA's for small or easy queries is a useful way of increasing effectiveness, but is problematic for larger queries or ones which are more complex. In "Web question answering: is more always better?" (p291-298) *Dumais et al* pointed out that redundancy is a resource in question and answering (Q&A) systems. It was asserted that Q&A is hard on limited data and that you have a better chance of finding an answer to a question if you have more data available. They use a simple technique of string-based re-writes for a query and apply an answer filter to the results. Their results using these simple techniques are comparable with other systems that use linguistic methods. The problem of predicting query performance without relevance information was addressed in "Predicting query performance" (p299-306) by *Cronen-Townsend et al*. They put forward a method called a 'clarity score' based on usage of words in queries compared to that of a collection. Queries which rank documents that are closely related to each other get a high clarity score, while queries that retrieve documents dealing with many more topics will achieve a lower score. They are able to show that there is a correlation between their 'clarity score' and average precision. The issue of query ambiguity was tackled by *Allen & Raghavan* in "Using part-of-speech patterns to reduce query ambiguity" (p307-314). Getting clarification from the user is

regarded a being important e.g. with the query 'fish', what is it about the fish concept which is important? Being able to fish, buy fish, eat fish etc. Parts of speech tagging can be used to generate these clarification questions. They are able to demonstrate practical benefits in the reduction in query ambiguity using this method.

Evaluation

In this session two papers were delivered on evaluation methods at TREC. *Voorhees & Buckley* in "The effect of topic set size on retrieval experiment error" (p316-p323) addressed the issue of the number of queries that need to be used in order to reduce error when comparing the performance of systems. They compute error rates for query sets up to the size of 25 and extrapolate in order to derive error rates on larger query sets. As the number of queries is increased for a set the error rate decreases, but at an asymptotic rate. It is argued that caution should be used when comparing two systems as error rates found in the study are larger than expected. If there is an error rate of 10% is it difficult to assert that system A is better than system B. Researchers should ensure that there are sufficient queries with the collection they are experimenting with to ensure that any comparisons have real value. *Sormumen* on the other hand looked at the issue of relevance judgements themselves in "Liberal relevance criteria of TREC – counting on negligible documents" (p324-330). It is argued that the criteria for accepting relevance are low and that more elaborate relevance schemes are needed. The use of graded relevance judgements is strongly argued for. He took 38 topics from TREC 7 and 8 together with a select number of documents (5737). Six assessors were employed in order to re-assess the documents regarded as relevant using graded relevance judgements. This paper came in for strong criticism from the audience as the assessors were shown documents that were known to have been judged relevant.

Efficiency

There were two papers in this session that concentrated on using compression techniques for fast query processing, while the other paper focused on a new approach for information retrieval. In "Efficiency phrase querying with an auxiliary index" (p215-221) *Bahle et al* described the use of an additional structure called the 'Nextword Index' in order to keep record of terms which following a given keyword. This is useful for servicing 'phrase based' queries, in many cases implicitly without the need for user action. For example if the user enters a query 'Presidents men', the system can use the 'Nextword Index' to retrieve records for the keyword 'President' which have 'men' adjacent. With compression, the amount of extra space needed is small (3% is quoted in the paper), with the cost of execution of phrase queries reduced by a factor of 3. Meanwhile *Scholer et al* in "Compression of inverted indexes for fast query evaluation" (p222-229) compared and contrasted bitwise and bytewise schemes for compressing records. Bitwise schemes stored compressed data (say integers) in terms of bits, while bytewise schemes store such in blocks. Their results show that a variable byte bytewise scheme allows query execution that is twice as fast at the bitwise scheme with only a small increase in storage space. The last paper in the session entitled "Set-based model: a new approach for information retrieval" (p230-237) by *Possas et al*, put forward a new

model for information retrieval based on the concept of 'termsets', rather than 'uniterms' used in current models. This leads to a new ranking mechanism for documents: using evidence from multiple terms rather than one term. They were able to demonstrate on the TREC-3 collection that their model yields an increase in retrieval effectiveness that is nearly double, but with an 80% increase in time for query execution.

Multimedia

There were two talks in this short session one on Music retrieval and another on video retrieval. The paper delivered by Shalev-Shwartz entitled "Robust temporal and spectral modelling for query by melody" (p331-338) put forward a method of query processing for Music IR, independent of tempo, dynamics, expression, accompaniment, instrument etc. They employ a strong statistical framework based on variable tempo in their system. It was found that as the duration of the query increases their model increase in performance, outperforming any fixed tempo model. They were able to show an average precision of around 95% and also gave a rather entertaining demonstration of their system. In "Video retrieval using an MPEC-7 based inference model" (p339-346), Graves and Lalmas focused on the structural, conceptual and contextual issues in video for the purposes of retrieval. In this way they demonstrate flexibility in query processing for video, for example in varying granularity of shot for ranking purposes or use context in order to constrain queries.

Acknowledgements

This report is funded by the John Cambell Trust, who kindly gave the author a grant to attend SIGIR in Tampere. The author also very grateful to both Ayse Goker and John Tait for attending the Collaborative filtering and Summarization parallel sessions respectively and giving me the feedback to write about those sessions.

References

[1] Beaulieu, M., Baeza-Yates, R., Myaeng, S.M and Jarvelin, K (eds), Proceedings of the twenty-fifth annual international ACM SIGIR conference on research and development in information retrieval, August 11 – 15, 2000, Tampere, Finland, ACM Press, 2002.

[2] Hawking, D., Overview of the TREC-9 Web Track. In: Voorhees, E.M., and Harman, D.K., (eds). The Ninth Text REtrieval Conference (TREC-9), NIST Special Publication 500-249, p87-102. [Available on http://trec.nist.gov/].