REPORT
on
ACM SIGIR WORKSHOP ON MATHEMATICAL/FORMAL METHODS IN INFORMATION RETRIEVAL
MF/IR 2002
Tampere, Finland
August 12-15, 2002

Sandor Dominich
Mounia Lalmas
C. J. "Keith" van Rijsbergen

Introduction

The previous two MF/IR workshops (ACM SIGIR 2000 MF/IR 2000, Athens, Greece, and ACM SIGIR 2001 MF/IR 2001, New Orleans, USA) showed that the mathematical/formal results achieved in Information Retrieval (IR) could be organized into a coherent theoretical framework, that they brought new knowledge to IR, and that mathematical/formal research in IR can stand as a specialized research area of IR. Therefore the purpose of the MF/IR 2002, which was the third in row, workshop was, on the one hand, to continue and enhance the results obtained so far, and on the other hand, to present, discuss, analyze, integrate the newer/newest results. MF/IR 2002 also aimed at promoting discussion and interaction among those with theoretical and applicative research interests in mathematical/formal aspects of Information Retrieval, and also at being a forum for the presentation of both theoretical and applicative results (e.g., foundational issues; description and/or integration of models; retrieval applications; mathematical/formal techniques, properties and structures in IR; existing and/or new theories and theoretical aspects).

The following areas were addressed: Information Retrieval, Information Filtering, Information, Mining, Indexing and Retrieval, Hypermedia, World Wide Web Retrieval, Digital Libraries, Evaluation, Human Computer Interaction, User Modeling, Semantic Web and Ontologies, where the different

entities involved (e.g., documents, queries, indexing, retrieval, relevance, effectiveness, users, etc.) were modeled using any of, but not necessarily limited to, the following approaches: Classical Sets, Fuzzy Sets, Rough Sets, Vectors, Linear Space, Similarity Functions, Probability, Theory of Uncertainty, Functional Analysis, Algebra, Topology, Metric Spaces, Euclidean Geometry, Non-Euclidean Geometries, Boolean Logic, Non-standard Logics, Fuzzy Logic, Quantum Logic, Matroid Theory, Graph Theory.

Papers

The event began with an opening talk given by C.J. van Rijsbergen entitled "The geometry of information retrieval". Van Risjbergen argued that basic concepts of quantum mechanics (states, observables, measurement) can be used to build a consistent theoretical — geometrical — framework for information retrieval. Probability can be derived from the geometry of the space, retrieval can be formally modeled using projection whereas relevance as an operator. Whether practical applications can be derived, or feedback and entanglement modeling can be treated within this geometric view are open questions.

The event continued with papers presentation as follows.

S. Robertson (Microsoft Research Cambridge, UK): *On Bayesian models and event spaces in information retrieval*.

Robertson discussed the issue of event spaces in probabilistic models. After arguing that the basic relationship to compute conditional probabilities can only be applied correctly to events that belong to the same probability space, Robertson concluded that the Robertson and Sparck-Jones [1975] model, denoted briefly by RSJ was not equivalent with the simple language model as claimed by Lafferty and Zhai (LZ) in a recent paper. The event space of RSJ is the space of documents, but it is not clear what should be taken as event space in the language model. The LZ model operates with cross-products of texts of queries and documents, not with events.

B. Thorsten (Palo Alto Research Center, USA): *Test data likelihood for PLSA models*.

The Probabilistic Latent Semantic Analysis (PLSA) model is fitted to a training corpus by an Expectation Maximization (EM) algorithm, which lacks the ability to handle unseen documents correctly and does not treat the undefined numeric case null is divided by null. Thorsten proposed an enhanced EM algorithm to solve this problem, and tested it on the MED collection with superior results.

Lee, C., and Lee, G.G. (Pohang Univresity, Korea): *Probabilistic information retrieval model for dependency structured indexing system*.

The Authors propose a method to incorporate term dependences into the probabilistic model. A structural indexing system is adapted to the 2-Poisson model, which consists of a dependency parse tree and Chow expansion. Experiments performed on the ETRI-KEMONG collection in Korean. Both the documents and queries were parsed and terms were extracted. Performance was evaluated by average precision. The results obtained were modest.

Dominich, S. (University of Veszprem, Hungary): *Paradox-free formal foundation of vector space model*.

The view that documents and queries belong to the same linear space yields well-known paradoxes. To overcome them, Dominich suggests that documents should form a probability space, the query is not included. As a noteworthy property it can be shown that the similarity measures reduce the quantity of the Shannon information assigned to it. Experiments are reported to illustrate this.

Goth, J. (University of Veszprem, Hungary): *Hyperbolic information retrieval*.

Goth investigates the possibility to use the Cayley-Klein hyperbolic geometry in the vector space model, and defines the hyperbolic information retrieval model (HIR). It is shown that the usual cosine measure and the new hyperbolic measure preserve the ranking of documents, which is illustrated with experiments.

Nie, J-Y, and Jin, F. (University of Montreal, Canada): *Integrating logical operators in query expansion in vector space model.*

Nie and Jin proposed that query expansion in the vector space model should mean the inclusion of terms into the original query that are alternative expressions of the original terms rather than additional ones. This can be achieved by OR-ing the expansion terms with the query terms. The expanded query is evaluated using a fuzzy set based definition of OR. The method was tested on the AP collection with good results.

Chang, Y., Choi, I., Choi, J., Kim, M., (Ajou University, Korea), and Raghavan, V.V. (University of Louisiana at Lafayette, USA: *Conceptual retrieval based on feature clustering of documents*.

The Authors propose a method to construct query concepts using the documents. Features are identified and extracted from documents using summarization techniques in vector format, they are then clustered into primitive concepts. The top ten most similar concepts to the initial query are used to generate DNFs, of which the one most similar to the query is linearly combined with the query; this is the used for retrieval. Experiments are reported using TREC-1 with good results.

Plachouras, V., and Ounis, I. (University of Glasgow, UK): *Query-biased combination of evidence on the Web*.

Plachouras and Ounis propose a query enhancement technique. A query scope is defined as a function of the scope of its terms using Wordnet which is viewed in two different ways: as a lattice and as independent terms. These two views are combined using Dempster-Shafer's theory of evidence. Experiments are reported using TREC10, they show that the proposed method doe snot outperform previous ones.

Hill, S.I., Zaragoza, H. (University of Cambridge, UK), Herbrich, R., and Rayner, P.J.W. (Microsoft Research Cambridge,UK): *Average precision and the problem of generalization.*

The Authors suggest a bound for average precision based on McDiarmid bounds, and show that in this way the deviation of observed

precision from its true value can be estimated. Thus, they argue, a learning theory-based treatment of information retrieval can be built.

Joachims, T. (Cornell University, USA): *Evaluating retrieval performance using clickthrough data*.

Joachims suggests a practical method, which is also theoretically analyzed, to evaluate Web retrieval functions using clickthrough data. Experiments were conducted using Google, MSNSearch and a Default strategy, the results are promising but still need analysis.

Friburger, N., and Maurel, D. (Informatics Laboratory Tours, France): *Textual similarity based on proper names*.

Friburger and Maurel propose a method to extract proper names from texts using a dictionary and automaton model. A piece of text is represented as a vector of its words and as a vector of its proper names (the IDF weighting scheme is used). The similarity between two texts is defined as a linear combination of the similarities (dot product) of corresponding words vectors and proper names vectors. Experiments are reported using the AMARYLLIS collection. Extensive experiments are reported and discussed.

Kim, H. (Electronics and Telecommunication Research Institute, Korea): *The influence of choice of record field on retrieval performance for bibliographic database*.

Kim reports on experiments conducted to investigate how the choice of record fields influence retrieval performance. Query terms were generated automatically using the INSPEC bibliographic and different fields separately: abstract, anywhere, descriptor, identifier, subject, title. Performance was evaluated using the D measure. The experiments showed that perfromance is sensitive to which field is selected, and the 'title' field yielded the best results.

Discussion, conclusion

Due to the high number of papers, but in order to allow for maximum interaction, each presentation was scheduled for 20 minutes including

questions. A 35 minutes discussion allowed further interaction between participants.

The following more important questions, ideas and views were raised (the number in brackets indicates the sequence of talk)

- Can a link be made between probability theory and possibility theory as regards the probabilistic model? [2]
- On what basis can we made the assumption that $\Sigma p < 1$? [3]
- The formulation of the vector space model as an IDO (information decreasing operator) was found to be a nice idea. [5]
- The use of fuzzy sets theory based formulas needs more justification, also their effect on rare and common terms. [7]
- More should have been said about the scaling up of the method proposed. [8]
- More experimentation is needed to test the generalization proposed. [10]
- More should have been said about the experiments (What articles have been clustered? What was the size of the collection used?) [12]

The MF/IR 2001 organisers were: Sandor Dominich (University of Veszprem, Hungary), Mounia Lalmas (Queen Mary, University of London, England, U.K.), and Keith van Rijsbergen (University of Glasgow, Scotland, U.K.), who, on behalf of MF/IR 2002 would like to thank the program committee for their help and time, and all authors and participants for writing and presenting papers as well as attending this event. They also would like to thank ACM SIGIR for making this event possible, and the University of Tampere and all local organizers for hosting it.