Information Retrieval, 8, 175–179, 2005

Guest Editorial: Spaces, Logic, and Link Analysis in IR: Recent Advances From A Mathematical and Logical Perspective

SÁNDOR DOMINICH dominich@dcs.vein.hu
Department of Computer Science, University of Veszprém, 8200 Veszprém, Egyetem u. 10, Hungary

MOUNIA LALMAS mounia@dcs.qmw.ac.uk
Department of Computer Science, Queen Mary University of London, England, United Kingdom

CORNELIUS JOOST (KEITH) VAN RIJSBERGEN keith@dcs.gla.ac.uk
Department of Computing Science, University of Glasgow, Scotland, United Kingdom

1. Introduction

The goal of this special issue is to present some of the theoretical and experimental research about aspects of modelling uncertainty, and about applying mathematical and logical techniques to Information Retrieval (IR). The papers in this special issue are, partly, the enhanced and thoroughly reviewed versions of the best presentations at the ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval, held in Tampere, Finland, in 2002, whereas the rest of them have been selected, through the usual reviewing process, from submissions in response to a call for papers for this special issue.

The first mathematical concept of a space, namely that of a linear space, was introduced into IR probably by Salton and his co-workers, when they defined the vector space model—for background see (Salton 1971). Although this concept was used more like a metaphor, Bollmann-Sdorra and Raghavan (1993) showed that using linear spaces generated some difficulties for basic concepts of IR, but at the same time their work drew attention to the fact that the concept of a space, provided it is used with enough care, is able to render structural properties of basic IR entities: these being document, query, relevance (Egghe and Rousseau 1998, Bollmann-Sdorra and Raghavan 1998, Dominich, 1999, 2001). More recently a unified approach to representing basic IR concepts in a linear space is achieved by modelling basic IR concepts using a Hilbert space in a Quantum Mechanical manner (van Rijsbergen 2004). Another type of space, that of a probability space, has also been applied to IR. Probabilistic techniques have been applied to IR as early as forty years ago (Maron and Kuhns 1960), and the probabilistic model for IR already took its known shape more than twenty years ago (Robertson and Sparck Jones 1976, van Rijsbergen 1979).

In this issue, several papers address, in a very exciting manner, the topic of applying probability and geometrical spaces and their techniques to IR: Are the probabilistic and language models different from one another? What probability spaces are being used in

IR? How many of them are there? If a lattice is associated with such a space, how can computational complexity be reduced? Is it possible to define retrieval in a non-Euclidean space? And, if possible, what are the benefits?

Perhaps the first application of rigorous logical techniques to IR can be traced back to the relational database model (Cooper 1964, Maron 1967), which inspired the exact technique, as we know it today, for the Boolean model. But the power of logic got its real personality in IR with the "logical uncertainty principle" associated with the view according to which retrieval can/should be conceived as an inference process (van Rijsbergen 1986, Crestani and van Rijsbergen 1995). This view has given birth to a new direction of research and opened up new perspectives generating both theoretical and practical research (van Rijsbergen and Lalmas 1996, Fuhr and Roelleke 1998, Lalmas 1998, He et al. 2002). In this issue, several questions are addressed, such as: How can logical tools be applied to Latent Semantic Analysis? How can they be applied to the combination of evidence in Web retrieval?

In Web retrieval, one of the underlying methods is the so-called link analysis. The starting point is the PageRank method, which stems from citation analysis (Geller 1978). This is concerned with the study of citation in the scientific literature. The underlying ideas—which are well known (and have a tremendous literature)—of citation analysis depend on citation counts as a measure of importance (Garfield 1955, 1972). In the PageRank method (Page et al. 1998), this approach is refined in that citation counts are not absolute values anymore, rather relative ones and mutually dependent. Other methods within link analysis are HITS and SALSA. In this issue, several important questions are being addressed: Will these algorithms always give a result? (Are they always convergent?) Are they stable, i.e., able to resist small changes in the input?

2. Papers in this issue

Thorsten Brants: Test data likelihood for PLSA models. In this nicely written paper, after a brief description of Probabilistic Latent Semantic Analysis (PLSA), the author points out that a disadvantage of the PLSA method is that it cannot assign probabilities to unseen test documents, and hence it assigns zeros to them. The paper proposes a new method to remedy this drawback, by which probabilities can be computed for unseen documents as well, these probabilities are maximal at the same points at which the true ones would be. The method proposed is based on folding-in, and makes it possible to calculate a likelihood for a test collection based on a model parameter. In order to test the retrieval performance (recall/precision at breakeven points) of the method as well as to compare it with that of other methods (likelihood based on marginalizing, word unigram, and partial prediction), experiments are carried out using the test collections MED and Reuters. Based on the results obtained, the author concludes that the proposed method based on folding-in is best suited to determine the number of iterations in an unsupervised learning setting.

Vassilis Plachouras, and Iadh Ounis: Experiments with query-biased combination of evidence on the Web. This is a meticulously composed paper. The authors begin by introducing the notion of query specificity, and define it as being a function of the specificity

of its terms. Two methods are proposed for the computation of query specificity using WordNet; one is based on viewing WordNet as a lattice, the other on conceiving WordNet as a collection of independent terms. Query specificity is then used in deriving a formula, based on Dempster-Shafer theory of evidence, to combine the content information and link structure information of the retrieved set of documents, as two bodies of evidence. In order to evaluate the "query-biased" combination of these two evidences, experiments are carried out using the WT10 g and GOV Web test collections. For content retrieval, the Divergence From Randomness framework is used, whereas for link analysis purposes two methods are used: the PageRank method and one called the Absorbing Model (an earlier method proposed by the authors). The results, which are extensively discussed, compare well with the baseline results, and even higher at places.

Maristella Agosti, and Luca Pretto: A theoretical study of Kleinberg's HITS algorithm. This is an elegantly written mathematical paper. After a compact presentation of the mathematical notions and results using graphs, matrices, eigenvectors, and Frobenius' theorem, a concise but very clear description of the HITS algorithm, and properties of a link graph are established. The authors reveal conditions when a link graph, such as that constructed for the Web or used in bibliometrics, has a dominant eigenvalue. They also prove that the HITS algorithm will always converge; its convergence has just been conjectured so far, but not proved. These theoretical results are then used to give a proper explanation to the practical behaviour of this algorithm in a real Web setting. Also, the authors suggest that a special version of the HITS algorithm can be applied to word stemming (Bacchin et al. 2005).

Ronny Lempel, and Shlomo Moran: Rank-stability and rank-similarity of link-based Web-ranking algorithms and authority-connected graphs. This is an exciting paper concerned with another aspect of link-based retrieval algorithms, namely that of stability (their 'ability' to resist small changes in their input) as regards to the scores they compute and rank order of hits they produce. The authors first give a brief overview of the PageRank, HITS and SALSA algorithms. Then, they present the mathematical concepts and results used in the paper. The authors give mathematical proofs that neither PageRank nor HITS is rank stable, whereas SALSA is. The importance of these results rests on an explanation as to why and how these algorithms are or not vulnerable to spamming.

Júlia Góth, and Adrienn Skrop: Varying retrieval categoricity using Hyperbolic Geometry. This is an interesting paper for two reasons. The first one is that it is concerned with a topic less well studied in IR, but which, however, can be important for users, namely the categorisation of answers (beyond their rank order), i.e., to what extent the Retrieval Status Values of the returned documents spread from each other. The less they do the more uncertain the user could be in which order to view the answers. The second reason is that the authors treat the subject by using a non-Euclidean Geometry: Cayley-Klein Hyperbolic Geometry. They propose a retrieval model based on this geometry, and then show that (i) it is equivalent to the Cosine based vector space model, and (ii) the proposed model allows for constructing retrieval systems that allow for varying the categorisation of answers at much lower computation costs than in the vector space model.

Karen SK Cheung, and Douglas Vogel: Complexity reduction in lattice-based information retrieval. This is an ingenious paper, and promising for practical applications. After a

brief presentation of the vector space model, concept lattices, and Singular Value Decomposition, it is shown how a term-by-document matrix can be associated with a concept-lattice. The authors propose a method, based on Singular Value Decomposition, to reduce the complexity, i.e., the number of elements, of this matrix.

Gloria Bordogna, and Gabriella Pasi: Personalised indexing and retrieval of heterogeneous structured documents. This is an interesting paper about personalised indexing. It proposes a method to index and retrieve structured documents based on fuzzy set theory. After a brief description of the mathematical concepts used (fuzzy membership function, cardinality, fuzzy binary relation, aggregation, linguistic quantifiers), a graph-based representation of documents is adopted as a hierarchical structure of sections, subsections, and paragraphs. An indexing method is then proposed consisting of two components: a static one, in which term weights are computed based on in which document part the term occurs, using fuzzy operators, and an adaptive component, in which personalised document representations are obtained using query terms. The authors also propose a retrieval method based on queries containing fuzzy operators.

Stephen Robertson: On event spaces and probabilistic models in information retrieval. This is an insightful paper concerned with understanding the relationship between probabilistic and language models of IR. Using examples, discussed at length and very elegantly, the author argues that the well-known formula from probability theory for conditional probability cannot be blindly applied. It is then made clear that, in the probabilistic model, the query and document belong to two different event spaces, and so do the query-document relevance pairs. This model assumes that there is one query for all documents; in other words, the event space of documents is re-constructed for every single query. In the language model, the probability of query is evaluated given a document. This would make the language model equivalent with the classical probabilistic model proposed by Maron and Kuhns (1960). Because this does not seem to be the original intention of the proponents of the language model, it is not clear what the event space is in this model: the document scores, for the same query, can hardly be compared with one another as they do not come from the same event space. The author concludes that the event space issue in both models has limitations, and awaits clarification.

Jacob Kogan, Marc Teboulle, and Charles Nicolas: Data driven similarity measures for K-means like clustering algorithms. The paper proposes a clustering algorithm using k-means. After a brief presentation of a basic k-means algorithm, an algorithm is proposed for clustering. The suggested new algorithm is described in great details; this is followed by a description of experiments carried out, using the Medlars, CISI, and Cranfield test collections, to evaluate it in terms of misclassified documents, and compare it with results obtained with other methods. The authors conclude that their results give best results for intermediate values of the parameters

3. Conclusion

The papers in this special issue address different aspects of several levels (logic, spaces, link analysis) of basic importance in IR. They demonstrate, once again, that the mathematical and logical methods in IR bring new theoretical and practical knowledge to IR.

References

Bacchin M, Ferro N and Melucci M (2005) A probabilistic model for stemmer generation. Information Processing and Management, 41(1):121–137.

Bollmann-Sdorra P and Raghavan VV (1993) On the delusiveness of adopting a common space for modelling IR objects: Are queries documents? Journal of the American Society for Information Science, 44(10):579–587.

Bollmann-Sdorra P and Raghavan VV (1998) On the necessity of term dependence in a query space for weighted retrieval. Journal of the American Society for Information Science, 49(13):1161–1168.

Cooper WS (1964) Fact retrieval and deductive question-answering information retrieval systems. Journal of the ACM, 11(2):117–137.

Crestani F and van Rijsbergen CJ (1995) Information retrieval by logical imaging. Journal of Documentation, 51(1):3–17.

Dominich S (1999) A geometrical view of relevance effectiveness in information retrieval. In Proceedings of Logical and Uncertainty Models for Information Systems. London, United Kingdom, pp. 12–22.

Dominich S (2001) Mathematical Foundations of Information Retrieval. Kluwer Academic Publishers, Dordrecht, Boston, London.

Egghe L and Rousseau R (1998) Topological aspects of information retrieval. Journal of the American Society for Information Science. 49(13):1144–1160.

Fuhr N and Roelleke T (1998) HySpirit—A probabilistic inference engine for hypermedia retrieval in large databases. In Proceedings of the 6th International conference on extending database technology (EDTB '98). Springer Verlag, pp. 24–38.

Garfield E (1955) Citation indexes for science: A new dimension in documentation through association of ideas. Science, 122(3159):108.

Garfield E (1972) Citation analysis as a tool in journal evaluation. Science, 178(4060):471–479.

Geller NL (1978) On the citation influence methodology of Pinski and Narin. Information Processing and Management, 14(2):93–95.

He D, Goker A and Harper DJ (2002) Combining evidence for automatic Web session identification. Information Processing and Management, 38(5):727–742.

Lalmas M (1998) Logical models in information retrieval: Introduction and overview. Information Processing and Management, 34(1):19–33.

Maron ME (1967) Relational data file I: Design philosophy. In: Schecter G, Ed., Information Retrieval: A Critical View, Thompson Book Company, pp. 211–223.

Maron ME and Kuhns JL (1960) On relevance, probabilistic indexing and information Retrieval. Journal of the ACM, 7(3):216–244.

Page L, Brin S, Motwani R and Winograd T (1998) The PageRank Citation Ranking: Bringing Order to the Web. Stanford University, http://dbpubs.stanford.edu:8090/pub/1998-66 (visited: 4 Nov 2002).

Robertson SE and Sparck Jones K (1976) Relevance weighting of search terms. Journal of the American Society for Information Science, 27(3):129–146.

Salton G (1971) Ed. The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall.

Van Rijsbergen CJ (1979) Information Retrieval. Butterworth, London.

Van Rijsbergen CJ (1986) A new theoretical framework for information retrieval. In Proceedings of the ACM Conference in research and Development in Information Retrieval, Pisa, pp. 194–200.

Van Rijsbergen CJ (2004) The Geometry of IR. Cambridge University Press.

Van Rijsbergen CJ and Lalmas M (1996) An information calculus for information retrieval. Journal of the American Society of Information Science, 47(5):385–398.