Improving formal models and usability: Research in Information Retrieval at Glasgow University

Editorial

Mark Dunlop and Mounia Lalmas

The search for a perfect information retrieval (IR) system can be compared to the search for the perfect blended whisky - taking inspiration from many components, maturing them together and finally satisfying the needs of a wide range of end-users. Like a blended whisky, IR is currently maturing together components from many constituent research communities, each with their own traditions and characteristics. On the formal side, these constituents include mathematical modelling of information using logical and probabilistic approaches, and modelling the information seeking process of searchers. These approaches are being added to a strong experimental and hypothesis testing tradition within IR research, which itself is being augmented by more psychological style experiments introduced to the computing science community via human computer interaction and cognitive science research.

The papers in this special issue present some of the research carried out in IR at the University of Glasgow, and highlight this mature blending of research approaches. Led by Professor Keith van Rijsbergen, the Glasgow IR Group has been and is investigating a wide range of research areas, as is reflected by the papers in this special issue. These areas range from formal approaches to the relevance and retrieval process, users and their information seeking behaviours, to the implementation and evaluation of these approaches.

The Glasgow IR group has carried out both theoretical and empirical work, aimed at giving end-users effective and efficient access to large collections of multimedia data. They have developed new models applying logics, uncertainty theories, and computational linguistics, to representing documents and queries, and to implementing retrieval functions. These models have been and are applied to large-scale experiments with text, image and speech data, as well as structured and hypermedia data. The Glasgow IR group is also involved in the development of interaction techniques that go beyond the traditional query and response paradigm, to include, for example, relevance feedback, case-based reasoning, and visualisation. The group is also researching on the evaluation of interactive IR systems, using methodologies adapted from human computer interaction and psychology. Many aspects of these research areas are reflected in the papers of this special issue. More information on the group can be found at http://ir.dcs.gla.ac.uk/.

This special issue of the *Journal of Information Retrieval* is composed of seven papers from researchers of the Glasgow IR group, ranging from theoretical models to evaluation methodologies.

The first paper, from *Keith van Rijsbergen*, entitled **Another look at the logical uncertainty principle**, proposes a formal mathematical approach for interpreting the IR process as a form of inference. The "logical uncertainty principle" was first advanced in [van Rijsbergen 1986] as a way of modelling the relevance of a document to a query as a probabilistic logical relation. This paper re-examines the principle from the point of view of classical logic. Keith van Rijsbergen is a professor of Computing Science at the University of Glasgow. He has been active in information retrieval research since 1969, covering both theoretical and experimental aspects. He has specified several theoretical models for IR and has seen some of them from the specification and prototype stage through to production. He has also contributed significantly to the experimental methodology for IR. He is author of the classic book *Information Retrieval* [van Rijsbergen 1979], and is the leader of the Glasgow IR group. His current research interests include dimensionality reduction, clustering, probabilistic retrieval [van Rijsbergen 1992], and logic-based information retrieval [Crestani et al 1995, van Rijsbergen 1993, van Rijsbergen and Lalmas 1996]. He has recently, together with Crestani and Lalmas, published a book entitled *Information Retrieval: Uncertainty and Logics* [Crestani, Lalmas and van Rijsbergen 1998].

The second paper, from *Fabio Crestani*, entitled **Exploiting the similarity of non-matching terms at retrieval time**, looks at the problem of "term mismatch". This is when a document is not retrieved in response to a query because document and query representations do not share any term, which is often the case with classic IR models. The paper investigates a new class of retrieval models that attempt to solve the problem by exploiting the availability of complete or partial knowledge of the similarity between terms. Fabio Crestani is currently a visiting research fellow at the International Computer Science Institute in Berkeley. Prior to this, he was a "Marie Curie" research fellow at the Department of Computing Science of the University of Glasgow and an assistant professor at the University of Padua, Italy. His research interests range from logical [Crestani, Lalmas and van Rijsbergen 1998] and artificial intelligence approaches to IR [Crestani and van Rijsbergen 1997], hypertext [Crestani and Melucci 1998], probabilistic IR [Crestani and van Rijsbergen 1998, Crestani et al 1998], to speech retrieval [Crestani et al 1997, Crestani 1999].

The third paper, from *Mark Sanderson*, entitled **Retrieving with good sense**, presents an analytical survey of the use of word sense disambiguation techniques to resolve ambiguity (for example, distinguishing the sporting sense of "bat" from the animal sense) with the aim of improving retrieval effectiveness. The paper discusses a number of approaches that were proposed as an attempt to solve word sense ambiguity for retrieval purposes. Mark Sanderson is now a lecturer in Information Science at the University of Sheffield. Prior to this, he was a post-doc researcher at the Center for Intelligent Information Retrieval, at the University of Massachusetts. He completed his thesis on disambiguation and IR in Glasgow in 1997 [Sanderson 1994, 1997] and works on summarisation [Sanderson 1998], hypertext retrieval [Harmandas, Sanderson and Dunlop 1997] and spoken document retrieval [Sanderson and Crestani 1998].

The fourth paper, from *Mirna Adriani*, entitled **Using statistical term similarity for sense disambiguation in cross-language information retrieval**, describes work on Indonesian to English and English to Indonesian cross-language retrieval. The issue under investigation is that of term ambiguity, which arises from the use of machine-readable bilingual dictionaries, used in cross-language retrieval. The paper proposes a sense disambiguation technique, which is based on a term-similarity measure for selecting the right translation sense for a query term. It also applies a query expansion technique based on the term similarity measure to improve retrieval effectiveness. Mirna Adriani is currently studying for a PhD in cross-language information retrieval [Adriani and van Rijsbergen 1999] and has done work involving a number of languages, in particular English, German, Indonesian, and Spanish. Before coming to Glasgow, Mirna was a visiting scholar in the Center for Intelligent Information Retrieval, University of Massachusetts.

The fifth paper, from *Ian Ruthven*, entitled **Interacting with characteristics of information use**, is a short paper that looks at some of the problems in interacting with ranked IR systems. After identifying the difficulty encountered by traditional IR systems in supporting information seeking, the paper proposes an alternative document representation, based on the use of information within documents. Ian Ruthven has been a research assistant in the Glasgow IR group since 1994 and is now studying for a PhD centred around developing formal methods to support interactive information retrieval [Ruthven, Lalmas and van Rijsbergen 1999, Ruthven and Lalmas 1999]. He is currently working on the project "Retrieval through explanation", funded by the British Library, together with van Rijsbergen and Lalmas. Prior to Glasgow, he gained a MSc in Cognitive Science from Birmingham and was a research associate in Medical Informatics at Manchester University.

The sixth paper, from *Iain Campbell*, entitled **Interactive evaluation of the Ostensive Model using a new test collection of images with multiple relevance assessments**, presents an interactive evaluation of different approaches to evidence combination within the "Ostensive Model". The paper also introduces the new image test-collection constructed for the evaluation. This model proposes a manner of structuring the uncertainty associated with individual relevance judgements as sources of evidence in relevance feedback. Iain Campbell is currently completing his PhD on the formalisation and exploitation of intuitions of user searching behaviour and building them into a query-less searching environment [Campbell and van Rijsbergen 1996]. He has been an information systems consultant for fifteen years. He has a particular interest in technology transfer - currently specialising in applying the theories of evidence combination and of information retrieval to commercial applications in message routing, user profiling, and knowledge management.

The final paper, from *Jane Reid*, entitled **A task-oriented non-interactive evaluation methodology for information retrieval systems**, describes a new evaluation methodology, using a task-oriented test collection, which combines the advantages of traditional non-interactive testing with a more user-centred emphasis. The methodology aims at capturing the many different types of relevance in information retrieval. The main features of the proposed methodology are the adoption of the task, rather than the query,

as the primary unit of evaluation, and the naturalistic character of the relevance judgments. Jane Reid studied modern languages at the University of St Andrews before obtaining a MSc in Information Technology at the University of Glasgow. She is currently completing her doctorate in evaluation of information retrieval systems [Reid 1999], and has recently taken up a post as a lecturer at Queen Mary and Westfield College, University of London. Her broader research interests include information seeking behaviours and other user-centred aspects of IR, and the application of human computer interaction techniques and principles to IR.

Biographies of the editors

Mark Dunlop is a Senior Scientist in the Centre for Human-Machine Interaction at Risø National Laboratory in Denmark. Prior to this post, he was a lecturer in human computer interaction and IR at the University of Glasgow. His research interests centre around usability of IR systems and IR techniques, and he currently works on evaluation of interactive IR systems [Dunlop, Johnson and Reid 1998; Dunlop 1997a], hypertext IR [Dalamagas and Dunlop 1997], relevance feedback [Dunlop 1997b] and information access on palm-top computers [Crossan and Dunlop 1999].

Mounia Lalmas is a lecturer at Queen Mary & Westfield College, University of London. In 1998, she was a part-time research assistant at the University of Dortmund working on the European projects EuroSearch and EuroGatherer, and a part-time research fellow at the University of Glasgow. Her research interests centre around the development of effective formalisms able to model information in the places and in the forms that it appears in an IR system. In particular, she is researching in the following areas: logical IR models [Lalmas and Bruza 1998], modelling uncertainty in IR [Crestani et al 98], modelling structured document indexing and retrieval [Lalmas and Ruthven 1998] and formal evaluations of IR systems [Lalmas and Ruthven 1999]. To maintain her successful collaboration with members of the Glasgow group, she has recently been appointed as an honorary lecturer at Glasgow.

The current Glasgow IR group

Since the deadline for papers for this special issue, Fabio Crestani has moved to the International Computer Science Institute in Berkeley, Mark Dunlop to Risø National Laboratory in Denmark, Mounia Lalmas and Jane Reid to Queen Mary & Westfield College in London, and Mark Sanderson to the University of Sheffield.

Matthew Chalmers and Joemon Jose have since joined the group as lecturers. Matthew Chalmers is working on social perceptual issues [Chalmers 1999], on both the system and theoretical side, in visualisation [Brodbeck et al 1997] and collaborative filtering [Chalmers, Rodden and Brodbeck1998]. Joemon Jose's research interests include multimedia information retrieval [Jose et al 1996], integration of databases and information retrieval, application of evidence combination techniques [Jose and Harper 1997], evaluation of information retrieval systems and object-oriented software IR architectures. Iadh Ounis will also shortly join the group as a lecturer – he is currently working at the National University of Singapore after completing a post-doc at CLIPS-

IMAG Grenoble. His research topics are logic-based information retrieval models, content-based multimedia retrieval, relational indexing and conceptual graph theory [Ounis 1998].

The Glasgow IR group also has several PhD students whose work was at too early a stage to report in this special issue, or who have joined the group recently. These include Di Cai working on evidential reasoning in modelling IR; Martin Gardner working on case-based reasoning, channel theory, hand-held computers, and design of context-sensitive interfaces and the Web; Marcos Theophylactou working on content based image retrieval and whose research interests also include the use of Dempster-Shafer's theory for natural language IR [Crestani et al 1997, Theophylactou and Lalmas 1998]; Tassos Tombros working on automatic text summarisation [Tombros and Sanderson 1998], document clustering, and presentation of retrieval results to IR users; and Robert Villa working on information representation, semiology, and user interaction with IR systems.

The Glasgow IR group has hosted many research visitors. One current visiting fellow is Gianni Amati, a researcher of the group Information Systems at the Fondazione Ugo Bordoni. He has collaborated with van Rijsbergen [Amati, van Rijsbergen, Ulbadini 1996] and Crestani [Amati, Crestani and Ubaldini 1997] on the probabilistic models of information retrieval and information filtering.

Acknowledgements

Selected bibliography

M. Adriani and C.J. van Rijsbergen. Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. *Proceedings of Research and Advanced Technology for Digital Libraries (ECDL)*, Paris, 1999.

G. Amati, F. Crestani, and F. Ubaldini. A Learning System for Selective Dissemination of Information. *Proceedings of International Joint Conference on Artificial Intelligence*, pp 764-769, Nagoya, Japan, 1997.

G. Amati, C.J. van Rijsbergen, and F. Ubaldini. The maximum expected utility principle and information retrieval. *Proceedings of the conference on information, statistics and induction in science*, pp 129 -140, Singapore, 1996.

D. Brodbeck, M. Chalmers, A. Lunzer & P. Cotture. Domesticating Bead: Adapting an Information Visualisation System to a Financial Institution. *Proceedings of IEEE Symposium on Information Visualisation,* Phoenix, pp 73-80, 1997.

I. Campbell and C.J. van Rijsbergen. The ostensive model of developing information needs. *Proceedings of the Second International Conference on Conceptions of Library and Information Science*, Copenhagen, pp 251-268, 1996.

M. Chalmers. Structuralist Informatics: Challenging Positivism in Information Systems. *Proceedings of Conference UK Academy for Information Systems,* pp 13-22, 1999.

M. Chalmers, K. Rodden & D. Brodbeck. The Order of Things: Activity-Centred Information Access. *Proceedings of WWW7,* Brisbane, pp 359-367, 1998.

F. Crestani, M. Lalmas and C.J. van Rijsbergen. *Information Retrieval: Uncertainty and Logics - Advanced models for the representation and retrieval of information*. Kluwer Academic Publishers, 1998.

F. Crestani, M. Lalmas, C.J. van Rijsbergen and I. Campbell. Is this document relevant? ... probably. A survey of probabilistic models in Information Retrieval. *Computing Surveys*, 30(4):528-552, 1998.

F. Crestani and M. Melucci. A case study of automatic authoring: from a Textbook to a Hyper-Textbook. *Data and Knowledge Engineering*, 27(1):1-30, 1998.

F. Crestani. Vocal Access to a Newspaper Archive: Design Issues and Preliminary Investigation. *Proceedings of ACM Conference on Digital Libraries*, pp 59-68, Berkeley, USA, 1999.

F. Crestani and C.J. van Rijsbergen. A study of probability kinematics in Information Retrieval. *ACM Transactions on Information Systems*, 16(3):225-255, 1998.

F. Crestani and C.J. van Rijsbergen. A Model for Adaptive Information Retrieval. *Journal of Intelligent Information Systems*, 8:29-56, 1997.

F. Crestani, I. Ruthven, M. Sanderson and C .J van Rijsbergen. The troubles with using a logical model of IR on a large collection of documents. Experimenting retrieval by logical imaging on TREC. *Proceedings of the Fourth Text REtrieval Conference (TREC-4), Washington D.C.*, 1995.

F. Crestani, M. Sanderson, M. Theophylactou, and M. Lalmas. Short Queries, Natural Language and Spoken Documents Retrieval: Experiments at Glasgow University. *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, Washington D.C., 1997.

A. Crossan and M.D. Dunlop. Dictionary based text entry method for mobile phones, *Proceedings of INTERACT 99,* Edinburgh, 1999.

T. Dalamagas and M.D. Dunlop. Automatic construction of news hypertext. *Proceedings Hypermedia - Information Retrieval - Multimedia (HIM)*, Dortmund, 1997.

M.D. Dunlop, C.W. Johnson and J. Reid. Exploring the layers of information retrieval evaluation, editorial for second special edition of *Interacting with Computers on HCI and IR*, 10(3): 225-236, 1998.

M.D. Dunlop. Time Relevance and Interaction Modelling for Information Retrieval, *Proceedings of the Twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, 1997a.

M.D. Dunlop. The effect of accessing non-matching documents on relevance feedback, *ACM Transactions on Information Systems*, 15(2), 1997b.

V. Harmandas, M. Sanderson and M. D. Dunlop. Image retrieval by hypertext links. *Proceedings of the Twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, 1997.

J.M. Jose and D.J. Harper. A Retrieval Mechanism for Semi-Structured Photographic Collections. *Proceedings. DEXA'97, LNCS 1308,* pp 276-292, Toulouse, 1997.

J.M. Jose, D.J. Harper, D.G. Hendry and J. Ijdens. A Spatial Feature Based Photograph Retrieval System. *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, pp 18-22, 1996.

M. Lalmas and P.D. Bruza. The use of Logic in Information Retrieval Modelling. *Knowledge Engineering Review*, 13(2):1-33, 1998.

M. Lalmas and I. Ruthven. Modelling the interaction in information retrieval. *Proceedings of the 9th European-Japanese Conference on Information Modelling and Knowledge Bases (EJ-IMKB)*, Iwate, Japan, 1999.

M. Lalmas and I. Ruthven. Representing and Retrieving Structured Documents using the Dempster-Shafer Theory of Evidence: Modelling and Evaluation. *Journal of Documentation*, 54(5):529-565, 1998.

I. Ounis, Organising Conceptual Graphs for Fast Knowledge Retrieval. Proceedings of International Conference on Tools with Artificial Intelligence, Taipei, Taiwan, pp 120-129, 1998.

J. Reid. A new, task-oriented paradigm for information retrieval: implications for evaluation of information retrieval systems. *Proceedings of the Third International Conference on Conceptions of Library and Information Science (CoLIS3)*, Dubrovnik, Croatia, pp 97-108, 1999.

I. Ruthven and M. Lalmas. Selective relevance feedback using term characteristics, *Proceedings of Third International Conference on Concepts in Library and Information Science (CoLIS3)*, Dubrovnik, Croatia, 1999.

I. Ruthven, M. Lalmas and C.J. van Rijsbergen. Retrieval through explanation: Inference approach to relevance feedback. *Proceedings of 10th Annual Irish Conference on Artificial Intelligence Cognitive Science (AICS)*, Cork, Ireland, 1999.

M. Sanderson. Accurate user directed summarisation from existing tools. In the *Proceedings of the 7th International Conference on Information and Knowledge Management* (CIKM 98), pp 45-51, 1998.

M. Sanderson. Word sense disambiguation and information retrieval. *Proceedings of the Seventieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 142-151, Dublin, 1994.

M. Sanderson. Word Sense *Disambiguation and Information Retrieval*. PhD Thesis, Department of Computing Science, University of Glasgow, 1997.

M. Sanderson and F. Crestani. Mixing and merging for Spoken Document Retrieval, *Proceedings of the 2nd European Conference on Digital Libraries*, Heraklion, Greece, pp 397-407, 1998.

M. Theophylactou and M. Lalmas. A Dempster-Shafer belief model for document retrieval using noun phrases. *20th Information Retrieval Colloquium British Computing Society: Discovering new worlds of IR*, Grenoble, 1998.

A. Tombros and M Sanderson. The advantages of query-biased summaries in Information Retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 2-10, 1998.

C. J. van Rijsbergen. The state of information retrieval: logic and information. *The Computer Bulletin*, 5(1):18-20, 1993.

C. J. van Rijsbergen. Probabilistic Retrieval Revisited. *The Computer Journal*, Special Issue, 35(3):291-298, 1992.

C. J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29:481-485, 1986.

C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979. (Available at http://www.dcs.glasgow.ac.uk/Keith/Preface.html).

C.J. van Rijsbergen and M. Lalmas. An Information Calculus for Information Retrieval. *Journal of the American Society of Information Science*, 47(5):385-398, 1996.