

ASIS&T ANNUAL MEETING AWARD WINNERS

A Career in Information Retrieval Research

by Kalervo Järvelin

ASIS&T 2012

Editor's Note: Each year that the ASIS&T Research Award is given we invite the recipient to share his or her research goals and discoveries with *Bulletin* readers. This year's recipient is Kalervo Järvelin, professor and vice chair at the School of Information Sciences, University of Tampere, Finland. He can be reached at Kalervo.Jarvelin@uta.fi.

EDITOR'S SUMMARY

Recalling his start in information science studies, 2012 ASIS&T Research Award winner Kalervo Järvelin explained that reading seminal books in the field influenced his academic and research path. A call to devise a curriculum for classification, indexing and information retrieval drew him away from computer science and firmed his career focus. Driven by the idea that information should be fully accessible for all, regardless of format, language or location, Järvelin pursued studies on the contribution of natural language processing, ontology-driven query expansion and feedback, cross-language information retrieval and metrics for retrieval evaluation. He spoke of challenges for information retrieval, including morphologic complexity of his native language and others, vocabulary mismatch between query and text, and optimal methods for assessing relevance of search returns. Järvelin's current research is on information retrieval in specific task settings and simulating human information behavior for more efficient analysis. An extensive list of publications and awards provides evidence of Järvelin's significant contributions to information science.

KEYWORDS

scholars
career development
information retrieval
multilingual retrieval
query processing
query expansion

Kalervo Järvelin, professor and vice chair at the School of Information Sciences, University of Tampere, Finland, is the 2012 recipient of the ASIS&T Research Award. He can be reached at Kalervo.Jarvelin@uta.fi.

Originally aiming to become a librarian, I was first introduced to information science and information retrieval (IR) by my first professor, Sinikka Koskiala, at the University of Tampere in 1972. She guided me to read F.W. Lancaster's *Information Retrieval Systems: Characteristics, Testing and Evaluation* (1968), Manfred Kochen's *The Growth of Knowledge* (1967), Gerard Salton's *Automatic Information Organization and Retrieval* (1968) and several other excellent texts. The ideas gained from them remained in my mind while I studied computer science, database management in particular, and nearly stayed in that area as a researcher. Accidentally, I returned to information science in the early 1980s and assumed the responsibility of developing the curriculum for classification, indexing and information retrieval at one of the predecessors of my current school, the School of Information Sciences, University of Tampere, Finland.

My initial research efforts were split between information seeking and knowledge-work augmentation on the one hand, and relational database management on the other. Both interests have continued to date, but IR has formed my main research area since the early 1990s. For the curious reader, my publications are listed at <http://people.uta.fi/~kalervo.jarvelin/KalPubl.html>, but most of them may be found through Google Scholar.

The initial driving aim in my research in IR was that all information should be available to anyone desiring it and in an accessible form, no matter in which form or language it is stored or where it is located. Today, much of this availability has been realized in the form of the web, its search engines and the resources accessible through them. With my colleagues in the research group FIRE, I have been happy to contribute to IR in the areas of natural language processing (NLP) method evaluation for IR, ontology-based query expansion and relevance feedback, cross-language IR (CLIR) methods/evaluation and IR evaluation metrics. This work has been great fun.

Originally, IR methods were developed for English, which is a morphologically simple language. This characteristic means that very simple methods of stemming are sufficient to make documents accessible as far as language is concerned. My native language is Finnish, which is highly inflectional. Every noun may have some 2000 inflectional forms, for example, in contrast to four forms in English. This complexity means that high recall is difficult to achieve with simple methods in Finnish – which has given Finnish benchmark-language status in NLP and CLIR experiments. Lemmatizers (see <http://en.wikipedia.org/wiki/Lemmatization>) were seen as necessary instead of stemmers for document representation. We also noted that many other languages are, while morphologically simpler than Finnish, clearly more complicated than English. We have not created stemmers or lemmatizers for any language ourselves but have evaluated their effectiveness for document representation in a range of languages. However, such tools cannot always be applied – if one has no control over database production – or be available at all for many languages. We have created lightweight statistical lemmatizers for indexing, and morphologically smart query-time tools for expanding the original query words to their most frequent inflectional forms of each language. We have shown such methods to be effective. These findings are good news in the global information access scene, where many languages are not nearly as well equipped as English.

One of the basic tough problems in IR is vocabulary mismatch: the searcher's query words do not match with the words in relevant documents. Ontology-based query expansion and relevance feedback are two approaches to solve the problem through query reformulation. Both expand the query with new words that are semantically, syntagmatically or (at least) statistically associated with the original ones and hopefully better match the relevant document texts. We were among the first to analyze the effectiveness of various query structures in semantic query expansion in best-match IR in late 1990s and identified the effective synonym structure for expansion. Interactive relevance feedback, while not really popular in practice among searchers, has been an appealing idea for query modification for a long time. Here the searcher examines the result of an initial query and identifies the relevant results for the search engine. We have shown through a number

of simulation studies that an effective approach is to provide feedback only on a few first results. This finding holds even if the first results are of marginal relevance and one aims to retrieve only highly relevant documents.

Cross-language IR methods gained in importance along with the global development of web IR. We developed in the late 1990s the dictionary-translation method for CLIR based on synonym structure. In a bilingual translation setting, the target language translation equivalents (for example, in English) of a single source language word (for example, Spanish) are all put into a synonym set in the target query without attempting to disambiguate various word senses. This simple method proved very effective and served as a challenging baseline in CLIR for a number of years. However, dictionary translation in CLIR may be bogged down by OOV (out-of-vocabulary) words. These words may be proper names spelled differently in different languages or technical terminology not covered by a machine-readable dictionary. We developed novel and effective approximate string-matching methods and statistical transliteration-based methods to overcome the OOV problems during 2000-2010.

My search engine is better than yours. – Statements like this one are often sought after in IR research and are based on IR evaluation, which is sometimes referred to as a hallmark and distinctive feature of IR research. In the early years of the U.S. National Institute of Standards' Text REtrieval Conference (TREC), in the 1990s, test-collection based evaluation used binary relevance assessments with a very liberal relevance criterion. In addition, the evaluation itself was dominated by a scenario where the (simulated) searcher was exhaustively searching for relevant documents. We asked the questions: What if most documents are of marginal value, others being highly relevant? What if early retrieval of a relevant document, of any degree of relevance, is far more valuable than late retrieval? These questions led to the development of evaluation methods by highly relevant documents and, in particular, to a family of evaluation metrics based on cumulated gain. Among the latter normalized discounted cumulated gain, the nDCG, became very popular in IR evaluation and also in operational development within search engine companies.

While the progress in the field of IR is astonishing and impossible for

JÄRVELIN, continued

anyone to follow-up in all detail, I currently believe that we can do a much better job in supporting information access in people's tasks and focused everyday-life information need situations. My current work focuses on task-based IR and interactive IR, including simulation of multiple query IR sessions.

Regarding task-based IR, we have collected comprehensive qualitative data in two task settings, research tasks in molecular medicine and administrative tasks in city administration. We are planning to continue these efforts in public administration and commercial companies. The data collection methods include interviewing, task performance shadowing, client-side interaction logging, photo logging through SenseCam and questionnaires. We have found how information needs in simple tasks are satisfied through one or a few organizational information systems, while complex tasks require a range of sources and traversing through several types of systems, not just one search engine. We also classified barriers in information access by their character (conceptual, syntactic and technological) and by their context of appearance (work task, system integration, or system) and analyzed how these depend on task complexity.

Taking the human searcher as an actor (and thus, a variable) in IR research design poses many challenges. Humans learn, get tired and are expensive to hire for experiments. At any step in an interaction, they may take a range of decisions that may lead to the termination of their search session with either success or frustration. Such decisions depend on many factors such as their personal traits, work task and search task, current situation in the search, search strategy, the quality of document representation or search platform among others. Human information access behavior can be modeled, to some degree, through behavioral probabilities observed in real life. This ability provides an opportunity to simulate interactive sessions in the computer economically and without (unprogrammed) learning effects or fatigue. In fact, one may run in reasonable time (hours) experiments involving many million interactive sessions and identify which kind of decisions or behaviors are likely to lead to successful results. We have recently shown that expected human fallibility in providing relevance feedback does not deteriorate search results and how important it is to consider time factors as opposed to plain ranking quality in IR evaluation. ■ *Publications on next page*

A Short Biography

Kal Järvelin (<http://people.uta.fi/~kalervo.jarvelin/>) is professor and vice chair at the School of Information Sciences, University of Tampere, Finland. He holds a PhD in information studies (1987) and two MSc degrees (library and information science, 1978, and computer science, 1983) from the same university where he started as a student in 1972. He was academy professor with the Academy of Finland in 2004–2009.

Kal Järvelin's research covers information seeking and retrieval, natural language processing and ontological methods in IR, IR evaluation and database management. He has co-authored over 250 scholarly publications and supervised 17 doctoral dissertations. Several of his former students have a recognized standing within information science. He has an H-index of 27 in Google Scholar and 13 in Web-of-Science (December 2012). He is particularly well cited for the work he has co-authored on IR evaluation methods, task-based information seeking and the integration of information seeking and retrieval research.

He has frequently served the ACM SIGIR conferences as a program committee member (1992-2009), conference chair (2002) and program co-chair (2004, 2006, 2014); and the ECIR, the ACM CIKM and many other conferences as program committee member. He was an associate editor of *Information Processing and Management* (2008-2012).

Kal Järvelin has received many awards beginning with the Finnish Computer Science Dissertation Award 1986 and continuing with several best paper awards with co-authors, including the ACM SIGIR 2000 Best Paper Award for the seminal paper on the discounted cumulated gain evaluation metric; the ECIR 2008 Best Paper Award for session-based IR evaluation; the IliX 2010 Best Paper Award for a study on task-based information access. He also received the Tony Kent Strix Award 2008 in recognition of contributions to the field of information retrieval, and, most recently, the ASIS&T Research Award 2012 in recognition of contributions to the field of information science.

Some Key Publications

1. Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (ACM TOIS)* 20(4), 422-446.
2. Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context*. Dordrecht, NL: Springer. ISBN 1-4020-3850-X
3. Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management* 31(2), 191 - 213.
4. Pirkola, A. & Hedlund, T. & Keskustalo, H., & Järvelin, K (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval* 4(3/4), 209-230
5. Kumpulainen, S., & Järvelin, K. (2010). Information interaction in molecular medicine: Integrated use of multiple channels. In: Belkin, N. et al. (Eds.). *Proceedings of the Information Interaction in Context Conference (IliX 2010), New Brunswick, NJ, August 2010* (pp. 95-104). New York, NY: ACM. ISBN: 978-1-4503-0247-0.
6. Keskustalo, H., Järvelin, K., Pirkola, A., Sharma, T., & Lykke Nielsen, M. (2009). Test collection-based IR evaluation needs extension toward sessions - A case of extremely short queries. In: Lee, G. & al. (Eds.). *Proceedings of AIRS 2009, the 5th Asia Information Retrieval Symposium, Sapporo, Japan, October 2009* (pp. 63-74). Heidelberg: Springer. (*Lecture Notes in Computer Science, 5839*).