Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval

W. Bruce Croft (editor)
(University of Massachusetts, Amherst)

Dordrecht: Kluwer Academic
Publishers (The Kluwer international
series on information retrieval, edited
by W. Bruce Croft), 2000, xv+306 pp;
hardbound, ISBN 0-7923-7812-1, $99.50,
£68.75, Dfl 230.00

Reviewed by
Sanda Harabagiu
Southern Methodist University

The recent advances in information retrieval (IR) in this collection of ten original papers reflect the wide range of research topics developed at the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts, Amherst. W. Bruce Croft, the Director of CIIR and the editor of this volume, presents in the preface both the history and the impressive research track of the center. The preface also lists the topics covered in the collection, reflecting the fact that the majority of the papers deal with research in traditional IR or with architecture and implementation issues. Only one of the papers tackles new areas, namely the topic detection and tracking problem. Surprisingly, none of the papers address one of the most exciting new IR tasks: the open-domain textual question-answering task. As one might expect, the papers in this volume are of varying quality. However, both the IR researcher and the computational linguist will find at least two of the papers outstanding.

Warren Greiff's paper "The use of exploratory data analysis in information retrieval research" reports on a new line of research that uncovers statistical regularities in novel ways. By analyzing data using the notion of weight of evidence, Greiff obtains a new formulation of the inverse document frequency (IDF) that generates results of the same quality as those obtained using traditional IDF measures. Moreover, this data-driven approach is extended to other IR measures, such as term frequency and document length. This approach, comprising four incremental models, generates a ranking formula that is shown to perform similarly to the INQUERY system. Greiff's data-driven model has special promise as it allows natural extensions based on additional sources of evidence such as thesaurus terms or phrases. This technique is an ideal research vehicle for traditional and modern IR.

The second paper that presents outstanding new research is "Topic detection and tracking: Event clustering as a basis for first story detection" by Ron Papka and James Allan. This paper captivates the reader by presenting an overview of the topic detection and tracking (TDT) problem, whose purpose is to organize broadcast news stories by the real-world events that they discuss. The paper describes the research problems considered in all three phases of the TDT and focuses on the solutions developed at CIIR for one of the problems, namely the *detection* problem—that is, identifying when new topics have appeared in the news stream. The approach discussed is both practical and technically interesting as it presents both new algorithms and modifications of existing IR techniques for the TDT problem.

Another strength of the volume comes from the expertise in language modeling developed at CIIR. The contribution of language models to several aspects of IR is well represented, appealing to researchers interested in statistical natural language processing (NLP). Two different language models are presented. The first model, originally introduced by Ponte and Croft (1998), is described from two different perspectives. In "Combining approaches to information retrieval," by Bruce Croft, the Ponte and Croft language model is considered when dealing with combinations of evidence generated by merging retrieval models and strategies. This paper also contains a wealth of information, representative of more than ten years of IR research into combining retrieval representations, retrieval algorithms, and search results. The same language model is also considered in "Language models for relevance feedback" by Jay Ponte, where relevance feedback and routing techniques are derived. Ponte's paper supports the theoretical findings with extensive experimental data.

A different, unigram language model is described in "Topic-based language models for distributed retrieval" by Jinxi Xu and Bruce Croft. This new language model, called *topic model*, is used to characterize the content of a specific topic from a given collection. In this paper, topics are approximated as document clusters, produced with the *k*-means clustering algorithm (Jain and Dubes 1988). This representation entails three new methods for distributed IR, suitable for different environments, including dynamic ones. Readers interested in other aspects of distributed IR are presented with an excellent account of the techniques involved in resource selection and merging of document rankings in Jamie Callan's paper "Distributed information retrieval." This paper lists experimental data that demonstrates the effectiveness of distributed IR techniques. The architecture of distributed IR systems is considered in "The effect of collection organization and query locality on information retrieval system performance" by Zhihong Lu and Kathryn McKinley. The authors' expertise in distributed and parallel systems ports new, interesting research perspectives to the problem of distributed IR.

The volume also contains papers on automatic derivations of concept hierarchies, cross-language retrieval and image retrieval. In "Building, testing, and applying concept hierarchies," Mark Sanderson and Dawn Lawrie present a method of automatically devising concept hierarchies based on document term frequencies, a metric that was recently used by Caraballo and Charniak (1999) to determine the specificity of nouns in texts. In "Cross-language retrieval via transitive translation," Lisa Ballesteros presents the CIIR efforts in cross-lingual IR, focusing on a dictionary-based approach. In "Appearance-based global similarity retrieval of images," S. Chandu Ravela and C. Luo present a technique of computing global appearance similarity, which enables appearance-based retrieval of images.

To the computational linguist, this collection of articles has interest for at least three reasons. First, the past years have shown that several NLP techniques play a central role in one of the most exciting new applications in IR: Open-domain textual question answering (Q/A)—that is, the task of producing answers in the form of text snippets from a corpus whenever an open-domain natural language question is posed.[1] The development of recent Q/A systems paved the way to new models of IR and especially to novel semantic processing mechanisms, able to operate on more and more complex questions and text passages. It is clear that this new interest in semantic processing will find its way into mainstream NLP.

1 Research in the area of information retrieval has been encouraged by the Text Retrieval Conference (TREC), sponsored by the National Institute of Standards and Technology. In 1999, TREC initiated a Question-Answering track, whose aim is to foster research in the area of textual Q/A, especially when pursued in a domain-independent manner.

Second, NLP approaches incorporating IR techniques or metrics are often practical solutions to difficult problems of text processing. Examples are the text segmentation method proposed by Hearst (1994, 1997), which employs the cosine similarity used in vector IR models; the dialogue manager described by Chu-Carroll and Carpenter (1998) using a routing module based on latent semantic indexing, a technique widely used in IR; and the more recent summarization methods of Berger and Mittal (2000) based on one of the new language models developed in IR (Ponte and Croft 1998).

Third, modern IR is both a provider of new language models, to be used in different natural language processing tasks, and an application in need of deeper semantic processing of language. In summary, the book makes very interesting reading for casual or serious developers of IR systems and anyone who is interested in obtaining pragmatic knowledge from text passages.

References

Berger, Adam and Vibhu Mittal. 2000. Query-relevant summarization using FAQs. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 294–301. Hong Kong.

Caraballo, Sharon and Eugene Charniak. 1999. Determining the specificity of nouns from texts. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP) and Very Large Corpora (VLC)*, pages 63–70.

Chu-Carroll, Jennifer and Bob Carpenter. 1999. Vector-based natural language call routing. *Computational Linguistics*, 25(3):361–388.

Hearst, Marti A. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16. Las Cruces, NM.

Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Jain, Anil K. and Richard C. Dubes. 1988. *Algorithms for Clustering Data*. Prentice Hall.

Ponte, Jay and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281.

Sanda Harabagiu is an Assistant Professor at Southern Methodist University. She received her Ph.D. from University of Southern California in 1997. Her research focuses on coreference resolution, question-answering, and modern IR. Harabagiu is the recipient of a NSF CAREER award. Harabagiu's address is Department of Computer Science and Engineering, Southern Methodist University, Dallas, TX 75275, e-mail: sanda@seas.smu.edu.