Summarization as a Means of Information Access: Utilizing Semantic Metadata

Eugenie Giesbrecht
FZI Research Center for Information Technology
Haid-und-Neu-Str. 10-14, 76131, Karlsruhe, Germany
www.fzi.de/ipe
giesbrecht@fzi.de

Abstract

The existing search engine interaction paradigm of typing in keywords and getting an enormous list of links is not suited for a lot of information seeking tasks. We see synthesis or summarization of information to satisfy users' information needs as an important step on the way to next generation information access systems. The idea is to explore the alternative geometrical models of meaning [14] based on the theory of Quantum Mechanics and Hilbert Spaces as a unifying framework for integrating semantic metadata into retrieval and summarization.

Keywords: Information Retrieval, Summarization, Question Answering, Geometrical Models of Meaning

1. MOTIVATION

A well-known paradigm of querying documents in a classical information retrieval system is by inputting keywords and matching them against the terms by which the documents are indexed. In reply, the user receives a list of links to be consulted. This is rather similar to searching, borrowing and looking for relevant information in books in the library [10]. Paradoxically, "information" retrieval has established itself as a pure document retrieval.

The so-called information overload causes the traditional library paradigm of information retrieval systems to be reconsidered. As discussed in [13, 10], the existing search engine interaction paradigm of typing in keywords and getting a list of results is not suited for a lot of information seeking tasks. Though in the meantime we are able to extract named entities and patterns in the textual information thanks to text analysis research, this kind of semantic metadata is still insufficiently integrated in information seeking technology. Having detected certain entities and relations in the text is not an ultimate goal. The next challenge is to make use of this information to satisfy the user's information needs. The vision is that information access systems should more directly answer our information needs by information extracted from the documents and as far as possible processed and synthesized into a coherent answer [10].

Thereby, synthesis or summarization of information as an answer to a user's query is an important step on the way to *cooperative* information access systems.

2. BACKGROUND AND RELATED WORK

This work is on the intersection of at least three distinct communities, where the research has been conducted more or less independently of each other - information retrieval, open domain question answering and summarization - and will draw, amongst others, upon the research from text mining and cognitive information processing. As a representational formalism, multi-dimensional geometrical models of meaning are being explored.

In the following, we briefly mention relevant work in corresponding disciplines and figure out the weaknesses and potential intersections of current approaches. Section 3 points at some of the research questions that emerge out of the latter.

2.1. Information Retrieval

At the Demo'08 panel session¹ discussing the future of the web, it has been emphasized by the representatives of the biggest search companies² that the search should become "task-centric" or "wish fulfilling". "The search engine should 'read' and synthesize the information to solve the intent", said Prabhakar Raghavan. The latter implies a fundamental change in the design assumptions of search systems by moving them from document search to the tasks for which people employ search. This is the idea behind this research - to synthesize information, presumably in a structured way, depending on the user's current intention.

2.2. Summarization and Open Domain Question Answering

Summarization is a rather new paradigm in information access and retrieval research. At the same time there is a long term tradition of research on summarization in NLP community, also specifically for purposes of open domain question answering. Automatic summarization here is a task of extracting the most important content from information sources and presenting it to the user in a condensed form and in a manner sensitive to the user's or application's needs [9].

The research in text summarization goes back to 1958, motivated by work of Luhn [8] who developed the first system of sentence extraction and building extractive summaries. The early work in open domain question answering goes back to the 90s. The ultimate goal of the latter has been to build systems that are able to answer any question in any domain. However, similarly to summarization and due to the inherent complexity of the task, this research focused mostly on extracting passages or sentences that have been ranked as most relevant to the question. The challenge here ended up in trying to get the most relevant sentence or passage as first-ranked. Consequently, the dominant approach in both summarization and question answering is still extraction rather than real abstraction, though both communities realize the need for abstraction and true synthesis of information [6, 11].

2.3. Dual Document Representations

The idea to move away from the *bag-of-words* in IR research, e.g. by mapping terms to concepts or accessing documents by extracted pieces of information, has been in the air for a while. Harris [5] already in 1959 proposed to extract certain relations from scientific articles by means of NLP and to use them for information finding. In order to achieve a kind of "conceptual" search, indexing strategies where the documents are indexed by concepts of WordNet [4], of Wikipedia [3] or ontology [1] have been used. It is just that the time is ripe now and information extraction technology has matured enough to use it on a large scale [10]. Most of these approaches have used either the one or the other way of "homogeneous" document representation. There are some attempts on the way to realize dual document representations by utilizing statistical language modelling (e.g. [2]). We are not aware, however, of any work aiming at integration of dual text representations by means of Hilbert Spaces.

3. RESEARCH DIRECTIONS

The goal of this thesis is to explore the ways to use the explicit semantic information, i.e. semantic metadata attached to the documents, not only to improve retrieval, but also to propose new ways of answering users' complex information needs in an "information overload" era in a cooperative way, i.e. by summarizing and allowing exploration based on the user's information need.

¹<http://www.demo.com/watchlisten/videolibrary.html?bcpid=1127798146bclid=1782597597bctid=1790936412>

²Peter Norvig from Google and Prabhakar Raghavan from Yahoo!

For this, the plan is to explore geometrical models of information retrieval [14] based on the theory of Quantum Mechanics and Hilbert Spaces, e.g. Tensor Space Models [7]. We believe, it is a promising alternative to leverage semantic metadata into the retrieval and summarization process.

Thereby, the suggested research is twofold:

1. to investigate the new geometry of IR and to exploit the possible ways of leveraging semantic metadata in the geometrical models of meaning and retrieval;
2. to explore the new ways of addressing users' information needs by means of summarization.

ACKNOWLEDGEMENTS

This work is supported by German Federal Ministry of Economics (BMWi) under the project THESEUS (number 01MQ07019).

REFERENCES

- [1] Bonino D., Corno F. (2008) Self-Similarity Metric for Index Pruning in Conceptual Vector Space Models. In *Proceedings of 19th International Conference on Database and Expert Systems Application (DEXA '08)*, pages: 225-229, Turin. IEEE Computer Society.
- [2] de Rijke M., Meij E., Trieschnigg D. and Kraaij M. (2008) Parsimonious concept modeling. In *31st Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*, July.
- [3] Gabrilovich E., Markovitch S. (2007) Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India.
- [4] Gonzalo J., Verdejo F., Chugur I, Cigarran J. (1998) Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the COLING/ACL 98 Workshop on Usage of WordNet for NLP*, pages 3844, Montreal, Canada.
- [5] Harris, Z. (1959). Linguistic transformations for information retrieval. In *Proceedings of the International Conference on Scientific Information*, Vol 2, Washington DC. National Academy of Sciences-National Research Council (NAS-NRC).
- [6] Lin J. (2007) Is question answering better than information retrieval? Towards a task-based evaluation framework for question series. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 212-219, Rochester, New York, April. Association for Computational Linguistics (ACL).
- [7] Liu N., Zhang B., Yan j., Chen Z., Liu W., Bai F., and Chien L. (2005) Text representation: From Vector to Tensor. In *Proceedings of ICDM*, pages 725-728. IEEE Computer Society.
- [8] Luhn, H. P. (1958) The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2, 157-165.
- [9] Mani I. (2001) *Automatic summarization*. John Benjamins Publishing Company.
- [10] Moens M.-F. (2006) *Information Extraction: Algorithms and Prospects in a Retrieval Context*. The Information Retrieval Series.
- [11] Nenkova A. (2008) Entity-driven rewrite for multidocument summarization. In *Proceedings of IJCNLP08*.
- [12] Nenkova A., Vanderwende L., McKeown K. (2006) A Compositional Context Sensitive Multidocument Summarizer: Exploring the Factors That Influence Summarization. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 573-580. Association for Computing Machinery (ACM).
- [13] Rose D. (2008) The information-seeking funnel. In *Position papers for the NSF Information Seeking Support Systems Workshop*.
- [14] van Rijsbergen, C. J. (2004) *The Geometry of Information Retrieval*. Cambridge University Press, August.