Applying semantically adapted vector space model to enhance information retrieval

Fisnik Dalipi[1], Ilia Ninka[2], Ajri Shej[3]

[1,3]Department of IT, Faculty of Math-Natural Sciences, Tetovo State University
fisnik.dalipi@unite.edu.mk, ajri.shej@gmail.com
[2]Department of IT, Faculty of Natural Sciences, University of Tirana
ilia.ninka@fshn.edu.al

Abstract. While most enterprise data is unstructured and file based, the need for access to structured data is increasing. In order to reduce the cost for finding information and achieve relevant results there is a need to build a very complex query which indeed is a serious challenge. Data volumes are growing at 60% annually and up to 80% of this data in any organization can be unstructured. In this paper we focus on describing the evolution of some modern ontology-based information retrieval systems. Further, we will provide a brief overview of the key advances in the field of semantic information retrieval from heterogeneous information sources, and a description of where the state-of-the-art is at in the field. Finally, we present and propose a novel use of semantic retrieval model based on the vector space model for the exploitation of KB (Knowledge Base) to enhance and support searching over robust and heterogeneous environments.

Keywords: ontology, information retrieval, semantic web, knowledge base.

1. Introduction

The phrase “information retrieval - IR” dates back to the 1950s [1], but the concept was firstly used at the library catalogues. Initial opinions on the subject emerged from librarianship and information science. Originally, this opinion had philosophical nature, dealing with how information should be classified and organized. Various schools held various positions and an ongoing debate was evident between them on philosophical and anecdotal rather than empirical grounds [2].

However, with the increasing volume of publication, and specifically of scientific literature, after the Second World War, practical concerns of how to effectively access this literature became urgent [3,4]. One of the most influential methods was described by H.P. Luhn in 1957, in which (put simply) he proposed using words as indexing units for documents and measuring word overlap as a criterion for retrieval [4]. In recent times, ontologies are widely used in IR systems.

Nevertheless, its main use has to do with query expansion, which consists in searching for the terms in the ontology more similar to the query terms, to use them together as a part of the query. In this work, we present and propose a novel use of semantic retrieval model based on the vector space model to enhance and support searching over robust and heterogeneous environments.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2. Semantic retrieval from heterogeneous environments

Semantic retrieval from distributed and heterogeneous environments is quite new concept and current ontology based retrieval technologies are very hypothetical, without having any well defined framework on applying ontology based search to the web as whole, which is consisted by unlimited number of domains. Some attempts have been made by [5], but they lack to address the potential use of ontology search beyond the organizational data corpus, as their models have difficulties to deal with the heterogeneity of Web and are limited to a predefined set of ontologies. The proposed architecture in Figure 3 reflects the concept of heterogeneity assuming large amount of semantic metadata online without having a pre-defined range of domains. We assume that the external element is not only a single knowledge base but involves online semantic web information.

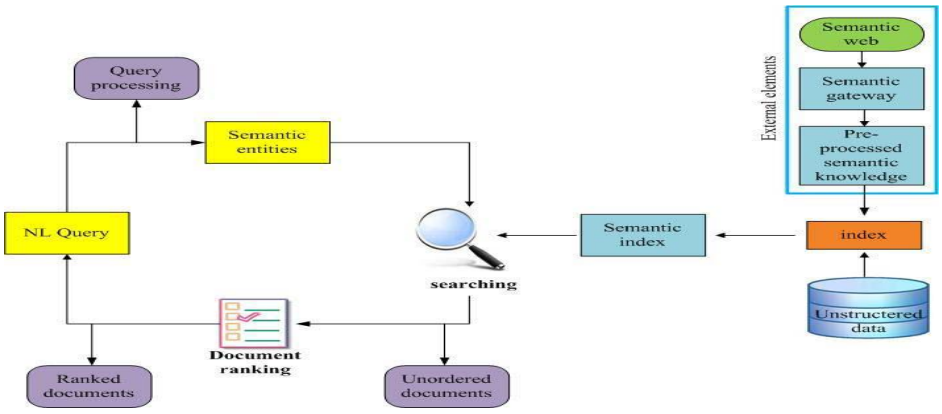

Fig. 3 Semantic information retrieval framework

This model does not require users to know special purpose query language; rather, the system expects queries to be expressed in natural language. Another relevant aspect is that the set of unstructured (web) information is not needed to be adapted into conventional fragments of ontological knowledge. In order to answer the queries, the system uses available semantic data and other information from standard web pages. When dealing with such a large amount of semantic information, we need a semantic gateway which will pre-process, gather, store and access the online distributed semantic web information. One of the most popular semantic way gateways currently available in the state of the art are: Watson [6], and Swoogle [7]. Once the user poses the query, that query can further be processed by any ontology based system which ensures access to the online ontologies and that translates generic natural language queries into SPARQL. Such systems of choice could be AquaLog, proposed by [8,9], Querix [10], or QASYO [11]. After returning the fragments of relevant ontological knowledge as an

answer, the system will perform a second step which includes retrieving and ranking by their probability the documents which contain the needed information. The ranking process can apply the concepts of vector space model ranking algorithm.

References

[1] S. Robertson. On the early history of evaluation in IR. In J. Tait, editor, Charting a New Course: Natural Language Processing and Information Retrieval – Essays in Honour of Karen Sp¨arck Jones, pages 13–22. Springer, 2005.

[2] S. Robertson. On the history of evaluation in IR. Journal of Information Science, 34 (4):439–456, 2008a

[3] W. Cleverdon. The significance of the Cranfield tests on index languages. In A. Bookstein, Y. Chiaramella, G. Salton, and V. V. Raghavan, editors, Proc. 14th Annual International ACM SIGIR Conference on Research and Development in In-formation Retrieval, pages 3–12, Chicago, Illinois, USA, Oct. 1991.

[4] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1(4):309–317, 1957.

[5] Maedche, A., Staab, S., Stojanovic, N., Studer, R., & Sure, Y. (2003). SEmantic portAL: The SEAL Approach. Spinning the Semantic Web. MIT Press , 317-359

[6] D'Aquin, M., Gridinoc, L., Sabou, M., Angeletou, S., & Motta, E. (2007). Characterizing
Knowledge on the Semantic Web with Watson. 5th International EON Workshop at International Semantic Web Conference (ISWC'07). Busan, Korea.

[7] Ding, L., Finin, T., Joshi, A., Pan, R., & Cost, S. (2004). Swoogle: A Search and Metadata Engine for the Semantic Web. 13th Conference on Information and Knowledge Management (CIKM 2004), (pp. 625-659). Washington, DC, USA.

[8] V. Lopez, M. Pasin, and Enrico Motta, "AquaLog: An Ontology-Portable Question Answering System for the Semantic Web," Lecture Notes in Computer Science, Vol. 3532, Springer, Berlin, pp. 546-562, 2005.

[9] V. Lopez, and E. Motta, "Ontology-Driven Question Answering in AquaLog," Lecture Notes in Computer Science, Vol. 3136. Springer-Verlag, Berlin, pp. 89–102, 2004.

[10] E. Kaufmann, A. Bernstein, and R. Zumstein., "Querix: A natural language interface to query ontologies based on clarification dialogs," In proceeding 5th International Semantic Web Conference (ISWC 2006), pp 980–981, 2006.

[11] A. M. Moussa and R. F. Abdel-Kader. QASYO: A Question Answering System for YAGO Ontology. International Journal of Database Theory and Application Vol. 4, No. 2, June, 2011

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288