Semantic information retrieval from heterogeneous environments

Fisnik Dalipi[1], Ilia Ninka[2]
[1]Department of IT, Faculty of Math-Natural Sciences, Tetovo State University
e-mail: *fisnik.dalipi@unite.edu.mk*
[2]Department of IT, Faculty of Natural Sciences, University of Tirana
e-mail: *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.

Index Terms— ontology, information retrieval, vector space model, semantic web, metadata, semantic index, 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].

There are two foundational projects that mark the creation of the information retrieval field. The Cranfield testing was the first project, and consisted of several experiments starting at 1957 and lasted until 1966. The second project was SMART, starting in the early 1960s and running in various forms until the end of the twentieth century. The emphasis of SMART has always been on purely automatic text retrieval - starting from an arbitrary piece of natural language text from the user and matching against automatically indexed documents [5].

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]. The algorithms developed in IR were the first ones to be employed for searching the World Wide Web from 1996 to 1998.

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.

2 IR concepts and models

IR represents a core component of the information systems. An information system must ensure that all the users who are meant to be served has the information needed to accomplish tasks, solve problems, and make decisions, no matter where that information is available. An information system must (1) actively find out what are the user's requirements or needs, (2) find and access documents, which results in a collection, and (3) match or affiliate documents with those requirements or needs. Realizing what type of information the user really needs to solve a problem is essential for successful retrieval.

The final goal of an IR system can be described as the representation, storage, organization of, and access to information items [6]. A global, abstract view of these elements is displayed in Figure 1.

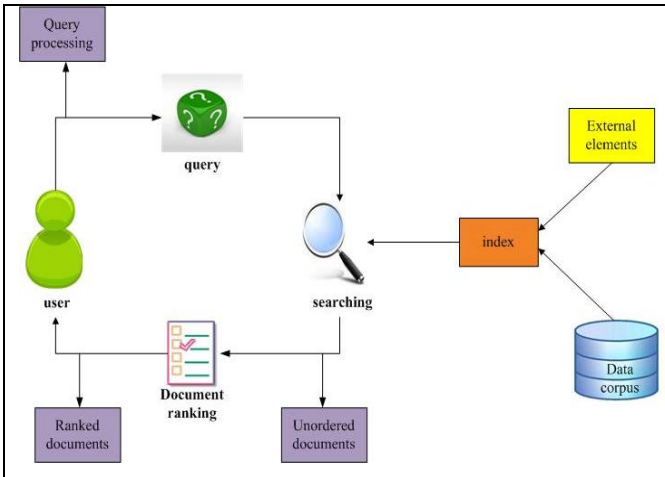

Fig. 1 Schematic view of the Information Retrieval process

The user formulates a query to the search engine. The format of the query can be search string, image, sound, etc. The search en-

gine then retrieves documents directly from the data corpus, but only those that are important to the query. External elements include the user interface, various query processing operations and resources for indexing (thesauri or controlled vocabularies). Next step is to rank them by decreasing probability or degree of relevance, and returns them to the user. This process can be iterated.
According to [7], we can distinguish four processes in an IR system: indexing, query processing, searching and ranking. The ranking step aims to predict which how relevant the items are comparatively to each other, thus returning them by decreasing order of estimated relevance. Thus, in a way, ranking algorithms can be considered the core of IR systems, as they are a key to determine the performance of the system.
Depending on the type of query, different mechanisms can be used to refine it. The most common ones are based on additional user input. In this spectrum, relevance feedback approaches are generally the most efficient ones. However, they reduce the usability of the systems, and therefore other external resources, such as taxonomies and thesauri, are often used instead (or complementarily) to automatically classify, disambiguate or expand query terms.

2.1 IR models and evaluation

The motivation for entering requests into an information retrieval system is an information need [8] and the success of a given retrieval system depends on the system's capacity to provide the user with the information needed with a reasonable time and with a straightforward interface for posing requests and collecting the results [9].
According to the definition in [6] an IR model is a quadruple [D,Q,F, sim], where:

- D is a set of (logical representations of) documents.
- Q is a set of (logical representations of) queries.
- F is a framework for modeling documents, queries, and their relationships.
- sim: $Q \times D \rightarrow U$ is a ranking function that defines an association between queries and documents, where U is a totally ordered set (commonly [0, 1], or P, or a subset thereof). This ranking and the total order in U define an order in the set of documents, for a fixed query.

There are standard measures to evaluate the performance of IR systems [10].
Precision: The ratio of documents retrieved by the system that are actually relevant to the query divided by the total number of documents retrieved.

$$P = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

For instance, if the system retrieved 6 documents for a query, where 3 of them were actually relevant, the precision performance for the system in that query is 0.5 or 50%. Polysemy may produce low precision rates, because irrelevant documents might be retrieved.
Recall: There may be many documents in the database that the user considers relevant, but only some of them will be retrieved by the system. The recall performance of a query is the number of relevant documents retrieved by the system divided by the total number of relevant documents in the database.

$$R = \frac{\text{Number of relevant documents retrieved}}{\text{Total of relevant documents in the collection}}$$

Response time: The elapsed time between the submission of a query and the presentation of the documents retrieved by the system. Precision could be easily maximized by retrieving a single document that is certainly relevant, and recall by retrieving all documents in the database. Thus, a measure that combines both of them is preferred, for example, the F-measure:

$$F = 2\frac{R\ P}{R + P}$$

where F-measure is the harmonic mean of precision and recall. The advantage of using F-measure is that maximizing it means maximizing a combination of recall and precision.
Three of the most important classic text IR models are: Boolean model, Vector space model and Probabilistic model.

2.2 Vector Space model

In the vector space model text is represented by a vector of terms [11]. The definition of a term is not inherent in the model, but terms are typically words and phrases. If words are chosen as terms, then every word in the vocabulary becomes an independent dimension in a very high dimensional vector space. Any text can then be represented by a vector in this high dimensional space. If a term belongs to a text, it gets a non-zero value in the text-vector along the dimension corresponding to the term. Since any text contains a limited set of terms (the vocabulary can be millions of terms), most text vectors are very sparse. Most vector based systems operate in the positive quadrant of the vector space, i.e., no term is assigned a negative value [12].
Following the presented previous notation, we define the vector space model as below:

- D: documents are represented by a vector of words or index terms occurring in the document. Each term in the document – or, for that matter, each pair (t_i, d_j) – has a positive, non-binary associated weight $w_{i,j}$.
- Q: queries are represented as a vector of words or index terms occurring in the query. Each term in the query– or, for that matter, each pair (t_i, q) – has a positive, non-binary associated weight $w_{i,q}$.
- F is an algebraic model over vectors in a t-dimensional space.
  - sim estimates the degree of similarity of a document d_j to a query q as the correlation between the vectors d_j and q. This correlation can be quantified, for instance, by the cosine of the angle between the two vectors:

    $\text{sim}(q,d_j)=\cos(q,d_j)=(q \bullet d_j)/(|q| \bullet |d_j|)$.

From the formula, sim(q, d_j) varies from 0 to 1, i.e. 1.0 for identical vectors and 0.0 for orthogonal vectors. Thus, instead of attempting to predict whether a document is relevant or not, the vector model ranks the documents according to their degree of similarity to the query.

3 SEMANTIC INFORMATION RETRIEVAL

Recently, ontologies have been used in Information Retrieval to improve recall and precision [13]. Its principal use is related to query expansion, which consists in looking for the terms in the ontology more related to the query terms, to use them as a part of the query. Much ontology has been designed for the purposes of managing and extracting semantic knowledge from online literature and databases.

IR systems that use semantic technologies for enhancing different parts of IR are called semantic search systems. Searching for the online ontologies, fact extraction from the ontologies, question answering, filtering and ranking retrieved information are usually put under the wing of semantic search. The introduction of ontologies to move beyond the capabilities of current search technologies has been an often portrayed scenario in the area of semantic-based technologies since the late nineties [14].

Based on literature review, below we provide the categories of semantic search engines. [15,16].

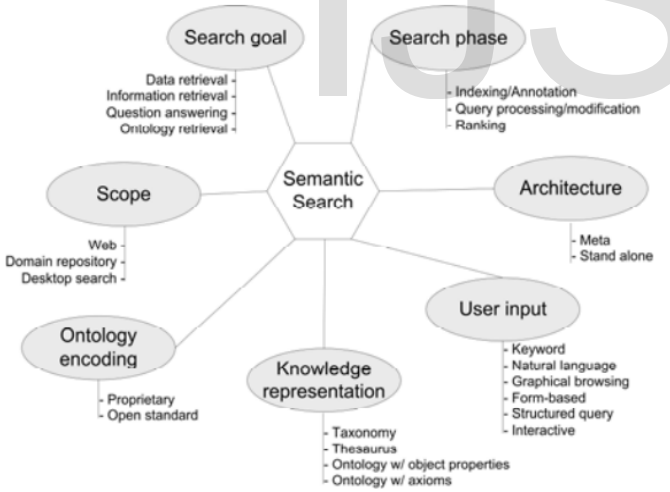

Fig 2. Classification of semantic search systems

In spite of the fact that that these concepts have shown several enhancements compared to classic keyword searching methodology, it is not clear among researchers that these techniques could be suitable to deal with large scale information sources.

4 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 [17], 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.

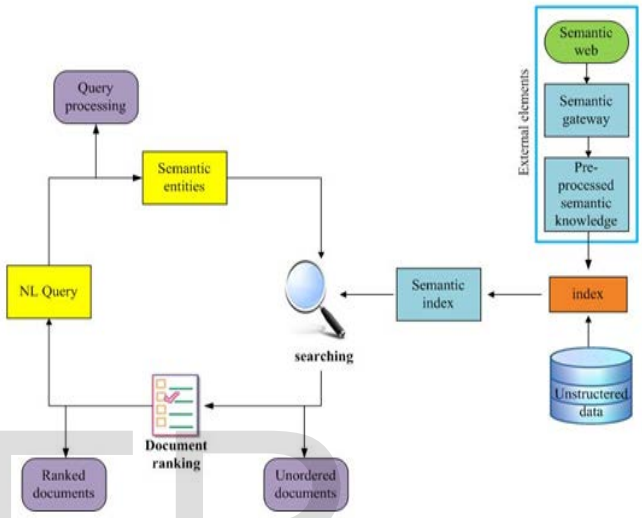

Fig 3. Semantic information retrieval framework

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 [18], and Swoogle [19].

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 Aqua-Log, proposed by [20,21], Querix [22], or QASYO [23].

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.

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 in-

volves online semantic web information.

5 Conclusions

In this paper, we focused on describing the evolution of modern IR systems with respect to their concepts and models and their application using ontologies. We define the IR evaluation and the vector space model. Further, we provided a brief overview of the key advances in the field of semantic information retrieval by describing where the state-of-the-art is at in the field.
Finally, we presented and proposed some ideas towards a novel use of semantic retrieval model which is based on the vector space model, aiming at enhancing and supporting the searching process over robust and heterogeneous environments with unlimited number of domains.

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] The S M A R T Information Retrieval Project, C. Buckley, G. Salton, J. Allan, Department of Computer Science , Cornell University, Ithaca , N Y, 14853. 1993 ACL Anthology

[6] Baeza Yates, R., & Ribeiro Neto, B. (1999). Modern Information Retrieval. Harlow, UK: Addison-Wesley.

[7] Croft, W. B., & Harper, D. J. (1993). Knowledge-based and statistical approaches to text retrieval. IEEE Expert: Intelligent Systems and their Applications , 8(2):8-12.

[8] Frederick Wilfried Lancaster. Information Retrieval Systems: Characteristics testing, Evaluation. John Wiley and Sons, second edition, 1979.

[9] Brice Austin. Mooer's Law: in and out of context. J.Am.Soc. Inf. Sci. Technol. 52(8): 607-609, 2001.

[10] Gerard Salton, A. Wong, and C. S. Yang. A vector space model for information retrieval. Communications of the ACM, 18(11):613–620, November 1975

[11] Van Rijsbergen. Information Retrieval Butterworths, 2nd edition, 1979. http://citeseer.ist.psu.edu/vanrijsbergen79information.html

[12] D. Harman. Information Retrieval: Data Structures and Algorithms, chapter
Ranking Algorithms, pages 363–392. Prentice-Hall, Englewood Cliffs, 1992

[13] T. Andreasen, J. Nilsson, and H. Thomsen.Ontology-based querying. In Proceedings of the Fourth International Conference on Flexible Query-Answering Systems, pages 15–26, Warsaw, Poland, Agosto 2000.

[14] Luke, S., Spector, L., & Rager, D. (1996). Ontology-Based Knowledge Discovery on the World-Wide Web. Internet-Based Information Systems: Papers from the AAAI Workshop. AAAI, (pp. 96-102). Menlo Park, California.

[15] Mangold, C. (2007) "A survey and classification of semantic search approaches", Int. J. Metadata, Semantics and Ontology, Vol. 2, No. 1, pp.23–34

[16] Esmaili, K.S. and Abolhassani, H. (2006) "A Categorization Scheme for Semantic Web Search Engines", In Proc. of IEEE Conf. on Computer Systems and Applications, pp 171-178.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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