ANALYZING SOME RECENT ARCHITECTURES FOR SEMANTIC SEARCH AT IR AND QA SYSTEMS THAT USE ONTOLOGIES

Fisnik DALIPI
Department of IT, Faculty of Math-Natural Sciences, Tetovo State University
Ilia NINKA
Department of IT, Faculty of Natural Sciences, University of Tirana

Abstract
Finding relevant information in the web, knowledge databases and different databases on personal computers is becoming a relevant task for companies. The quantity of information is increasing substantially including data from unstructured or structured resources. 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 will try to scan the current situation of IR (Information Retrieval) and QA (Query Answer) systems that use ontology. Further we'll analyze and discuss the process of reporting in visual search in multidimensional information spaces and we evaluate some architectures for semantic search combined with modern visual and interactive technics.
Keywords: Query, knowledge databases, Information Retrieval, Query Answer, reporting, visual selection.

1. Introduction

The methodology of archiving written information can be traced back to around 3000 BC, when the Sumerians designated special areas to store clay tablets with cuneiform inscriptions. Even then the Sumerians realized that proper organization and access to the archives was critical for efficient use of information. They developed special classifications to identify every tablet and its content. The need to store and retrieve written information became increasingly important over centuries, especially with inventions like paper and the printing press. Soon after computers were invented, people realized that they could be used for storing and mechanically retrieving large amounts of information [1]. 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 [2]. The algorithms developed in IR were the first ones to be employed for searching the World Wide Web from 1996 to 1998.

Using the query, an IR system retrieves information that might be relevant to the user. Question/Answering (QA) is a line of research in natural language processing, where a user poses a question in natural language, e.g. "Who is the winner of Nobel peace prize in 2011?" and expects an answer as in word/phrases or a sentence. Thus, QA research aims to deal with a wide variety of questions including: fact, list, definition, cross-lingual questions etc.

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.

2. Models and implementation of IR systems

IR represents a 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.

In the beginning, IR systems were boolean systems which allowed users to specify their information need using a complex combination of boolean ANDs, ORs and NOTs. Boolean systems have several shortcomings, e.g., there is no inherent notion of document ranking, and it is very hard for a user to form a good search request. However, most everyday users of IR systems expect IR systems to do ranked retrieval. IR systems rank documents by their estimation of the usefulness of a document for a user query. Most IR systems assign a numeric score to every document and rank documents by this score [3]. Several models have been proposed for this process. The three most used models in IR research are the vector space model, the probabilistic models, and the inference network model.

2.1 Vector space models

In the vector space model text is represented by a vector of terms [4]. 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 [5].

To assign a numeric score to a document for a query, the model measures the similarity between the query vector (since query is also just text and can be converted into a vector) and the document vector. Typically, the angle between two vectors is used as a measure of divergence between the vectors, and cosine of the angle is used as the numeric similarity (since cosine has the nice property that it is 1.0 for identical vectors and 0.0 for orthogonal vectors). If *D* is the document vector and *Q* is the query vector, then the similarity of document *D* to query *Q* (or score of D for Q) can be represented as

$$Sim(\vec{D},\vec{Q}) = \sum_{t_i \in Q,D} w_{t_{iQ}} \cdot w_{t_{iD}}$$

where w_{tiQ} is the value of the ith component in the query vector Q , and w_{tiD} is the ith component in the document vector *D* .

2.2 Probabilistic models

This family of IR models is based on the general principle that documents in a collection should be ranked by decreasing probability of their relevance to a query. This is often called the probabilistic ranking principle (PRP) [6]. Since true probabilities are not available to an IR system, probabilistic IR models estimate the probability of relevance of documents for a query.

2.3 Inference Network Model

In this model, document retrieval is modeled as an inference process in an inference network [7]. Most techniques used by IR systems can be implemented under this model. In the simplest implementation of this model, a document instantiates a term with certain strength, and the credit from multiple terms is accumulated given a query to compute the equivalent of a numeric score for the document. From an operational perspective, the strength of instantiation of a term for a document can be considered as the weight of the term in the document, and document ranking in the simplest form of this model becomes similar to ranking in the vector space model and the probabilistic models described above.

2.4 Implementation and evaluation of IR systems

Most operational IR systems are based on the inverted list data structure. This enables fast access to a list of documents that contain a term along with other information (for example, the weight of the term in each document, the relative positions of the term in each document, etc.). A typical inverted list may be stored as:

$$t_i \rightarrow < d_a ,\ldots >, < d_b ,\ldots >, \ldots , < d_n ,\ldots >,$$

which depicts that term-i is contained in d_a , d_b, . . . , d_n , and stores any other information. All models describe above can be implemented using inverted lists. Inverted lists exploit the fact that given a user query, most IR systems are only interested in scoring a small number of documents that contain some query term. This allows the system to only score documents that will have a non-zero numeric score. Most systems maintain the score for documents in a heap (or another similar data structure) and at the end of processing return the top scoring documents for a query. Since all documents are indexed by the terms they contain, the process of generating, building, and storing document representations is called indexing and the resulting inverted files are called the inverted index [5].

Objective evaluation of search effectiveness has been a key element of IR. There are standard measures to evaluate the performance of IR systems [8].

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 [26]

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

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.

3. Question/Answering systems

Many researches have been done in recent years on QA systems. QA systems have been expanded to answer simple questions correctly; but now researches have been focused on methods for answering complex questions truthfulness. Those methods analyze and parse complex question to multi simple questions and use existing techniques for answering them [9]. Recent researches show that increasing the performance of system is dependent on number of probable answers in documents. Finding the exact answer is one of the most important problems in QA systems. QA is a type of information retrieval. Given a collection of documents (such as the World Wide Web or a local collection), the system should be able to retrieve answers to questions posed in natural language. QA is regarded as requiring more complex natural language processing (NLP) techniques than other types of information retrieval such as document retrieval, and it is sometimes regarded as the next step beyond search engines [9]. QA research attempts to deal with a wide range of question types including: fact, list, definition, how, why, hypothetical, semantically-constrained and cross-lingual questions. Search collections vary from small local document collections to internal organization documents to complied newswire reports to the World Wide Web.

QA systems are classified in two main parts [10]: open domain QA system and closed domain QA system. Open domain question answering deals with questions about nearly everything and can only rely on general ontology [11] and world knowledge. On the other hand, these systems usually have much more data available from which to extract the answer. Closed-domain question answering deals with questions under a specific domain

(for example medicine or weather forecasting and etc) and can be seen as an easier task because NLP systems can exploit domain-specific knowledge frequently formalized in ontology.

4. IR and QA based on ontologies

Recently, ontologies have been used in Information Retrieval to improve recall and precision [12]. 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 apply semantic technologies to enhance different parts of IR are called semantic search systems. Searching for the online ontologies, fact extraction from the ontologies and question answering are usually put under the wing of semantic search.

There are two main directions when evaluating semantic search systems. First direction is to confirm their prevalence over existing search engines; and second – to evaluate its potential usage. Aiming at the first direction, the TREC[1] document collection would be a natural choice. However, this is still problematic since online ontologies cover only a little fraction of test collection's queries [13]. Consequently, most of Semantic IR evaluations are concentrated on their own data set. The second evaluation direction has not yet found its way into the community. It has been argued that information retrieval affects and is dependent on its context: task goals, information system, information used, information acquired and task process properties [14]. An important factor in search is the experience of users. Expertise in this area is often considered along two dimensions, namely, domain expertise and search expertise. The former subjects are knowledgeable about a particular domain, while the latter have experience in using search engines and tools. Domain experts evaluate search results more closely as well as web search experts investigate results deeply, while search novices use breadth-first search strategy [15].

5. Towards semantic-based information retrieval from heterogeneous information sources

Data spaces, large collections of heterogeneous data, and personal information management systems have recently received a lot of attention in the Database and Information Retrieval (IR) communities [16].

Retrieving information from distributed heterogeneous information sources represents a challenging problem. As the quantity of information is increasing substantially including data from unstructured or structured resources more intelligent retrieval techniques, focusing on information content and semantics, are required. Nowadays, we encounter the challenging issues of dealing with distributed and heterogeneous data sources containing huge amounts of data in varieties of semantic structures. Developing a data integration system is a complex undertaking which consists of major issues that may include the heterogeneity of the underlying data sources, the alteration in access mechanisms, and the support of query languages and approaches of semantic heterogeneity in relation to their data models. Recently, ontologies are being extensively applied to eliminate the problem of semantic heterogeneity. Here, we refer to some architectures which use ontologies for data integration to enable access to distributed heterogeneous data sources. Among these architectures, most notable mentioned in the research literature are: HeC[2], TAMBIS[3], SEMEDA[4] etc.

Each of these approaches use partially similar model including data warehousing with query which is based on ontology. The data warehousing approach uses single and centralized data storage to physically hold a copy of the data from each data source [17]. The model is presented below in Figure 1.

The schema in the data warehouse retains the collective schema of all data sources (called the global schema), and the ontology that is built on top of the global schema is called the global ontology. Here the schema defines the database at the logical level while the ontology defines the database at the conceptual level; mappings are provided between a schema and the ontology to link them. User queries are formulated on the global ontology and all requests are directly answerable by the warehouse. This can results in fast responses and enables multifaceted results from a centralized data store. Managing a data warehouses is also not a simple task. Whenever new data is added or removed from any of the source systems the update has to be reflected in the warehouse and this may require suspension of the execution of user data requests.

This architecture is often called an information push model, where the data is “pushed” into the data warehouse at definite times.

[1] Text REtrieval Conference - An annual information retrieval conference and competition, the purpose of which is to support and further research within the information retrieval field.

[2] Health-e-Child (HeC) project aims to develop an integrated healthcare platform for European pediatrics, providing seamless integration of traditional and emerging sources of biomedical information.

[3] The TAMBIS project was developed to provide transparent access across disparate biological databases with concepts specified using description logic based Ontology language, namely DAML+OIL.

[4] SEMEDA can be used to collaboratively edit and maintain ontologies, and to query the integrated databases in real time.

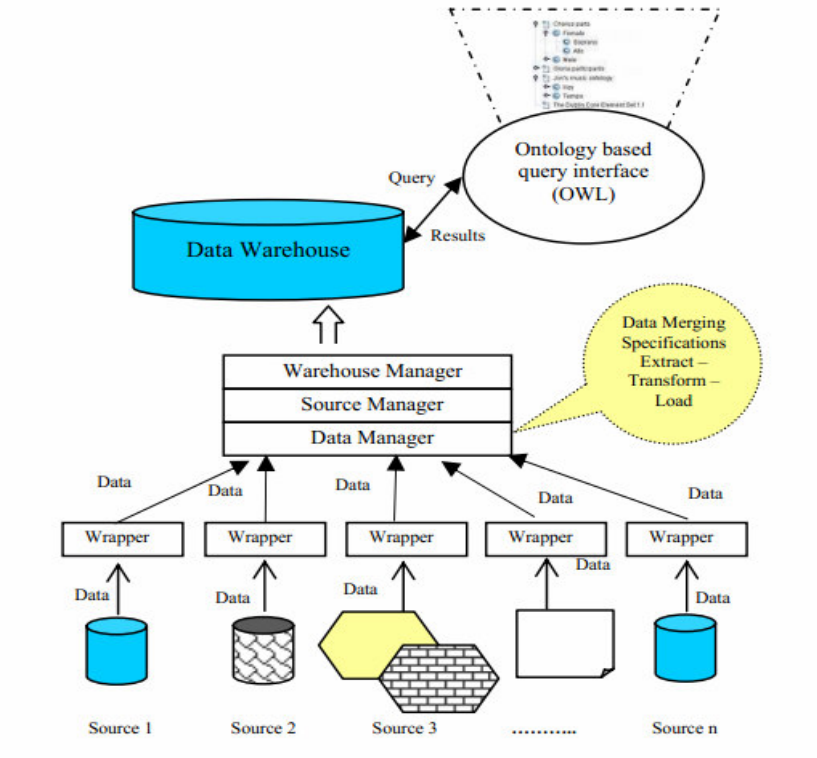

Figure 1.Data warehousing approach with ontology based query

6. Conclusion

In this paper, we focused on describing the evolution of modern IR and QA systems and their application using ontologies. Further, we provided a brief overview of the key advances in the field of semantic information retrieval from multidimensional heterogeneous information sources, and a description of where the state-of-the-art is at in the field.
Finally, we briefly explained the general data integration approach that utilizes ontologies to provide access to distribute heterogeneous data sources namely data warehouse and mediation approach.

7. References

[1] D. Harman. Information Retrieval: Data Structures and Algorithms, chapter

Ranking Algorithms, pages 363–392. Prentice-Hall, Englewood Cliffs, 1992

[2] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1957.
[3] J. Becker, and D. Kuropka, "Topic-based Vector Space Model", In Proceedings of the 6th International Conference on Business Information Systems, 2003
[4] [28] 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.
[5] Singhal Amit. Modern Information Retrieval: A brief overview. In IEEE Data Engineering Bulletin 24(4), pages 35-43, 2001.
[6] S. E. Robertson. The probabilistic ranking principle in IR. Journal of Documentation, 33:294–304, 1977.
[7] Howard Turtle. Inference Networks for Document Retrieval. Ph.D. thesis, Department of Computer Science, University of Massachusetts, Amherst, MA 01003, 1990. Available as COINS Technical Report 90-92
[8] M. Mauldin. Conceptual Information Retrieval:A case study in adaptive partial parsing. Kluwer Academic Publishers, 1991.
[9] Demner-Fushman, Dina, "Complex Question Answering Based on Semantic Domain Model of Clinical Medicine",OCLC's Experimental Thesis Catalog, College Park, Md.: University of Maryland (United States), 2006.
[10] Magnini, B., Negri, M., Prevete, R., Tanev, H.: "Is It the Right Answer? Exploiting Web Redundancy for Answer Validation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), Philadelphia, PA. 2002.

[11] Figueira, H. Martins, A. Mendes, A. Mendes, P. Pinto, C. Vidal, D ,"Priberam's Question Answering System in a Cross-Language Environment",LECTURE NOTES IN COMPUTER SCIENCE, Volume 4730, 2007,PP. 300-309
[12] 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.
[13] d'Aquin, M., Motta, E., Sabou, M., Angeletou, S., Gridinoc, L., Lopez, V., and Guidi, D. (2008) "Toward a new generation of semantic web applications", IEEE Intelligent Systems, Vol 23, No. 3, pp 20-28.
[14] Ingwersen, P. & Järvelin, K. (2005). The Turn: Integration of Information Seeking and Retrieval in Context. Heidelberg: Springer. ISBN 1-4020-3850-X. For more information on this book
[15] Jenkins, C., Corritore, C. and Wiedenbeck, S. (2003) "Patterns of information seeking on the web: A qualitative study of domain expertise and web expertise", IT and Society, Vol 1, No. 3, pp 64-89
[16] J. Gennari, M. Musen, R. Fergerson, W. Grosso,M. Crub´ezy, H. Eriksson, N. Noy, and S.Tu.The evolution of Prot´eg´e–2000: An environment for knowledge–based systems development. In-ternational Journal of Human-Computer Studies,58(1):89–123, 2003
[17] K.Munir, M.Odeh, R.McClatchey, S.Khan, I. Habib. Semantic Information Retrieval from Distributed Hetereogeneous Data Sources. Presented at the 4th International Workshop on Frontiers of Information Technology -- FIT 2006. Islamabad, Pakistan December 2006