UDC:004.65
Original scientific paper

A DOCUMENT RETRIEVAL METHOD BASED ON ONTOLOGY ASSOCIATIONS

Jan Paralič, Ivan Kostial
Department of Cybernetics and AI, Technical University of Kosice,
Letna 9, 040 11 Kosice, Slovakia
{Jan.Paralic, Ivan.Kostial}@tuke.sk

Abstract: *In this paper a new, ontology-based approach to information retrieval (IR) is presented. The system is based on a domain knowledge representation schema in form of ontology. New resources registered within the Webocrat system[1] are linked to concepts from this ontology. In such a way resources may be retrieved based on the associations and not only based on partial or exact term matching as the use of vector model presumes. In order to evaluate the quality of this retrieval mechanism, experiments to measure retrieval efficiency have been performed with well-known Cystic Fibrosis collection of medical scientific papers. The ontology-based retrieval mechanism has been compared with traditional full text search based on vector IR model as well as with the Latent Semantic Indexing method..*

Keywords: *information retrieval, document retrieval, ontology-based retrieval, vector representation, latent semantic indexing.*

1. INTRODUCTION

A considerable amount of explicit knowledge is scattered throughout various documents within organizations and people minds working there. In many cases the possibility to efficiently access (retrieve) and reuse this knowledge is limited [3]. As a result of this, most knowledge is not sufficiently exploited, shared and subsequently forgotten in relatively short time after it has been introduced to, invented/discovered within the organization. Therefore, in the approaching information society, it is vitally important for knowledge-intensive organizations to make the best use of information gathered from various information resources inside the organizations and from external sources like the Internet [7]. On the other hand, tacit knowledge of authors of the documents' provides important context to them, which cannot be effectively intercepted.

Knowledge management [8] generally deals with several activities relevant in knowledge life cycle [1]: identification, acquisition, development, dissemination (sharing), use and preservation of organization's knowledge. Our approach to knowledge management in the e-Government context supports most of the activities mentioned above.

[1] The Webocrat system has been developed within the EC funded project IST-1999-20364 Webocracy (Web Technologies Supporting Direct Participation in Democratic Processes)

Based on this approach, a Web-based system Webocrat [6] has been designed and implemented. It has been tested on pilot applications at Wolverhampton (UK) and in Košice (Slovakia). Firstly, it provides tools for capturing and updating of tacit knowledge connected with particular explicit knowledge inside documents. This is possible due to ontology model, which is used for representation of organization's domain knowledge. Ontology with syntax and semantic rules provides the 'language' by which Webocrat(-like) systems can interact at the *knowledge level* [5].

Use of an ontology enables to define concepts and relations representing knowledge about a particular document in domain specific terms. In order to express the contents of a document explicitly, it is necessary to create links (associations) between the document and relevant parts of a domain model, i.e. links to those elements of the domain model, which are relevant to the contents of the document.

Model elements can be also used for search and retrieval of relevant documents. In case all documents are linked to the same domain model, it is possible to calculate a similarity between documents using the abovementioned conceptual structure of this domain model. Such approach supports also 'soft' techniques, where a search engine can utilize the domain model to find concepts related to those specified by user. The search engine can thus return every document linked to the concepts, which are close enough to the concepts mentioned in the user's query.

In order to evaluate retrieval efficiency of such an ontology-based approach, we did a series of experiments with two other frequently used techniques for information retrieval (vector model with *tf-idf* weight schema and latent semantic indexing model). In the following section 2, all three retrieval methods are briefly described. Section 3 describes data set used for the experiments as well as the results achieved. Finally, section 4 provides a summary of the experimental results and suggestions for future work.

2. SCHEME OF DOCUMENT RETRIEVAL

We developed package with three different approaches to document retrieval: vector representation [2], latent semantic indexing method (LSI) [2], and ontology-based method used in the Webocrat system. In next sub-chapters, each of these approaches is briefly described.

2.1. VECTOR REPRESENTATION APPROACH

This well know approach is based on vector representation of document collection. First of all every document is passed through set of pre-processing tools.

1. *Lower case filter* transforms all the letters on their lower case;
2. *Stop words filter* eliminates all words provided on the so called stop list (these are less informative words such as articles, connectives etc.);
3. *Document frequency filter* eliminates all terms that occur more frequently than a given upper frequency threshold or less frequently than a lower frequency threshold.

Then a vector of index term weights is calculated as the document internal representation. These weights are calculated by most often used *tf-idf* scheme as follows [4]:

$$w_{ij} = tf_{ij} \times idf_i$$

$$\text{where } tf_{ij} = \frac{freq_{ij}}{\max_e freq_{ej}} \text{ and } idf_i = \log\left(\frac{N}{n_i}\right),$$

$freq_{ij}$ is the number of occurrences of term t_i in document d_j and $max_e freq_{ej}$ is the frequency of the most frequent term t_e within the document d_j. N is number of all documents in collection, and n_i is the document frequency for term t_i in the whole document collection (number of documents from collection where term t_i is presented).

Such a vector is then normalized to unit length and stored into the term-document matrix A, which is internal representation of the whole document collection.

In order to find some relevant document to a specific query $\vec{Q}$ it is necessary to represent the query $\vec{Q}$ in the same way as a document $\vec{D}_i$ (i.e. a vector of index term weights). Similarity between a query $\vec{Q}$ and a document $\vec{D}_i$ is computed as cosine of those two normalized vectors (document and query vectors).

$$sim_{TF-IDF}(\vec{Q}, \vec{D}_i) = \frac{\vec{D}_i \times \vec{Q}}{\left|\vec{D}_i\right|\left|\vec{Q}\right|}$$

2.2. LATENT SEMANTIC INDEXING APPROACH

LSI approach is based on singular value decomposition of tf-idf matrix A. By this decomposition three matrices are computed [9].

$$A = USV^T$$

where S is the diagonal matrix of *singular values* and U, V are matrices of left and right *singular vectors*. If the singular values in S are ordered by size, the first k largest values may be kept and the remaining smaller ones are set to zero. The product of the resulting matrices is a matrix approximately equal to A, and is closest to A in the least squares sense.

$$A \cong A_{SVD} \text{ where } A_{SVD} = U_K S_K V_K^T$$

In order to determine similarity between a query and approximate document vector $\vec{D}_{i,SVD}$, we need to transform query vector to new feature space. (Original query vector is computed with tf-idf scheme as described above for vector model approach.)

$$Q_{SVD} = Q_{TF-IDF}^T U_K S_K^{-1}$$

and then we can compute similarity in the same way as before, i.e.

$$sim_{SVD}(\vec{Q}_{SVD}, \vec{D}_{i,SVD}) = \frac{\vec{D}_{i,SVD} \times \vec{Q}_{SVD}}{\left|\vec{D}_{i,SVD}\right|\left|\vec{Q}_{SVD}\right|}.$$

2.3. ONTOLOGY-BASED APPROACH

This part describes the Webocrat-like approach that uses ontology for document retrieval purposes. For the experiments described below we did not consider type of relation in ontology for calculation of similarity between concepts. Moreover, we assumed that the set of relevant concepts to the query is known. But this condition can be achieved with any technique for assigning concepts from ontology to a query, e.g. based on manual

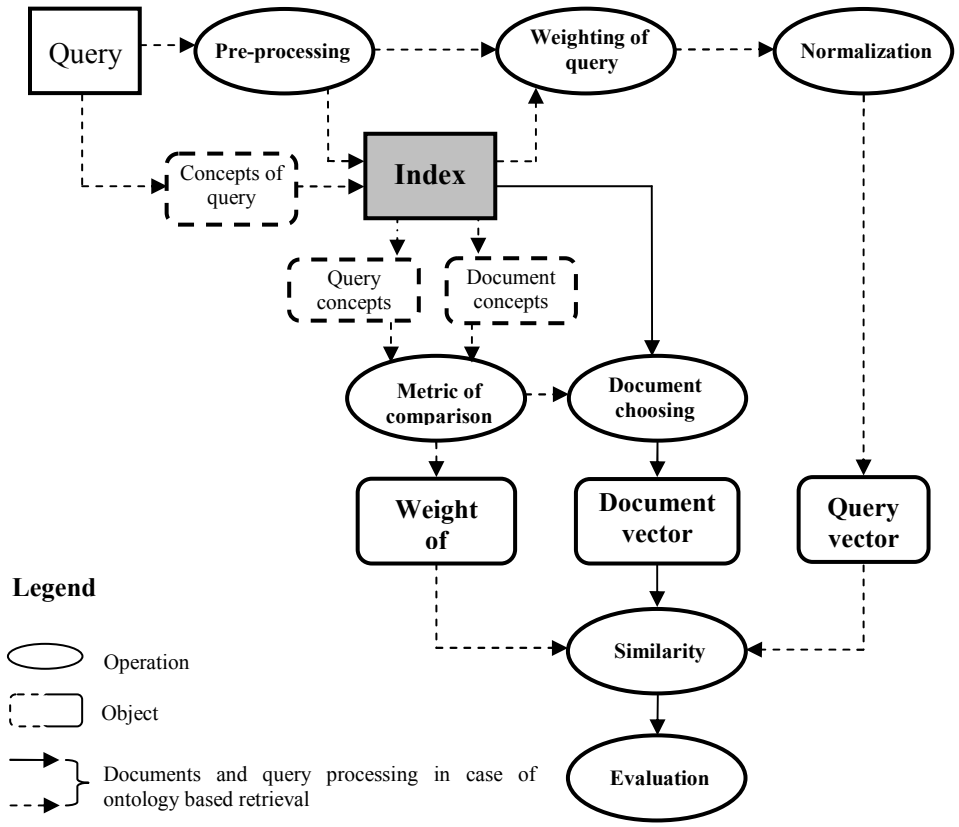

Figure 1: The process of ontology based document retrieval

assignment or based on synonyms to query terms, making use of Wordnet or other.

The way in which a query is processed by this approach is shown on the Figure 1. For a given query first appropriate concepts are retrieved - in our case manually from the user. Then the set of concepts associated with each document is retrieved from database. As next, these two sets are compared using simple metric, which expresses the similarity between a document $\vec{D}_i$ and given query $\vec{Q}$.

$$sim_{onto}(\vec{Q},\vec{D}_i) = \begin{cases} \left|Q_{con} \cup D_{i,con}\right| & if \quad \left|Q_{con} \cup D_{i,con}\right| \neq 0 \\ k \end{cases}$$

where Q_{con} is a set of concepts assigned to query $\vec{Q}$ and $D_{i,con}$ is a set of concepts assigned to document $\vec{D}_i$, and k is small constant, e.g. 0.1. Resulted number represents ontology-based similarity measure.

Better results have been achieved when this number have been combined with some of the previous two retrieval approaches described above (i.e. LSI approach or vector model). The final similarity is then computed as multiplication, e.g.

$$sim(\vec{Q},\vec{D}_i) \quad = sim_{onto}(\vec{Q},\vec{D}_i) * sim_{TF-IDF}(\vec{Q},\vec{D}_i)$$

3. EXPERIMENTS

3.1. DOCUMENT COLLECTION

Collection named Cystic Fibrosis was used for our experiments. This collection consists of 1239 files [1]. It is a subset extracted from a large MEDLINE collection where a keyword *Cystic Fibrosis* was used. The minimal size of a file is 0.12 kb, maximum size is 3.8 kb and average size is 1.045 kb. A file with 100 queries is also supplied with the document collection. A set of relevant documents to each query is also provided. Each document in the answer set is ranked with respect to its relevance to the query by more experts - and can take values from 0 to 8 – see

Table 1. In our experiments a document has been taken into account as relevant to a query, if its average experts ranking was more than 4.

It is possible to look at this collection as a group of documents and concepts of ontology, where every document is assigned to an appropriate set of concepts and similarly every concept can "hold" some documents. There are 821 concepts and average number of concepts assigned to a document is 2.8. Similarly we can refer to concepts of this collection by the same way. Average number of documents assigned to one concept is 4.2.

Table 1: Cystic fibrosis document collection

Name of collection	Relevance rank	Minimal number of documents	Maximal number of documents	Average number of documents
Cystic fibrosis	3	1	131	17,95
	4	1	121	15,59
	5	1	114	13,5
	6	1	96	11,03

3.2. COMPARISON OF VARIOUS APPROACHES FOR DOCUMENT RETRIEVAL

In this section we will describe comparison of document retrieval experiments, where 3 different approaches were used: full text search (vector representation approach), latent semantic indexing approach, and ontology based approach. First approach was used as described above with lower document frequency threshold equal to 0.2% and upper threshold set to 80%, i.e. only terms with documents frequency from the given interval have been taken into account for index. Threshold for LSI dimension reduction *k* was set to 100.

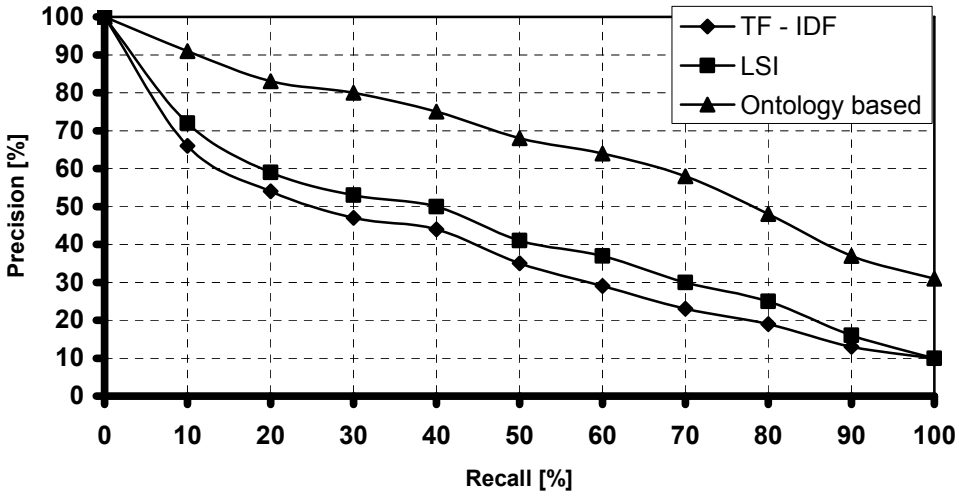

Figure 2: Precision-recall curve for three analyzed retrieval approaches

Precision-recall curve for all of the approaches described above are presented in Figure 2. Our experiments showed that the Webocrat-like approach based on an ontology is very promising, providing better retrieval efficiency than LSI or standard full text approach. However, as mentioned above, manual assignment of concepts to query has been used.

4. CONCLUSIONS

In this paper we have presented results of some experiments performed in order to evaluate document retrieval efficiency of an ontology based approach, which is implemented within the Webocrat system. We did a series of experiments with two other, frequently used techniques for information retrieval (vector model with tf-idf weight schema and latent semantic indexing model). The experiments on well-known Cystic Fibrosis document collection have shown that ontology based approach employed in the Webocrat system is very promising and may yield better precision-recall characteristics.

However, there are still open questions related to this approach. Probably the major one is the question how to transform a user-defined query into a set of concepts from actual ontology. In our approach this has been predefined (it could be inferred from given query set with proper answers that are available from Cystic Fibrosis collection). But this question may be definitely solved in many different ways. For example with any technique for assigning concepts from ontology to a query, e.g. based on manual assignment or based on synonyms to query terms, making use of Wordnet or other techniques.

Our future work will be focused on further enhancement of ontology-based retrieval mechanism using more sophisticated inference mechanism for finding similar concepts to given query. E.g. by analyzing different types of relations within actual ontology. There may be also other experiments with different combinations of analyzed approaches in various (real) settings etc.

ACKNOWLEDGEMENTS

This work has been done within the VEGA project 1/1060/04 ”Document classification and annotation for the Semantic web” of the Scientific Grant Agency of Ministry of Education of the Slovak Republic.

REFERENCES

[1] Abecker A., Bernardi A. Hinkelmann K. Kühn, O. & Sintek M. (1998): Toward a Technology for Organizational Memories, IEEE Intelligent Systems, 13, May/June, pp. 40-48.
[2] Berthier Ribeiro-Neto, Ricardo Baeza-Yates. *Modern Information Retrieval*. ACM Press, New York, 1999.
[3] Borghoff U. M. & Pareschi R. Eds. (1998) Information Technology for Knowledge Management. Springer Verlag
[4] S. Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, and Computers*, 23(2):229–236, 1991.
[5] Newell A. (1982) The Knowledge Level. Artificial Intelligence, 18, p. 87-127.
[6] Mach, M. – Sabol, T.: Knowledge-based System for Support of e-Democracy. In: Proc. of The European Conference on e-Government, Remenyi, D. – Bannister, F. (Eds). MCIL Reading, Dublin, September 2001, pp. 269-278
[7] Mach, M., Machova, K.: Knowledge Technologies for Information Acquisition and Retrieval. Proc. of the III. ISC'2003 – 3rd Internal Scientific Conference of the Faculty of Electrical Engineering and Informatics, Košice, 2003, pp. 61-62
[8] Tiwana A. (2000) The Knowledge Management Toolkit. Prentice Hall.
[9] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

Received: 17 December 2003
Accepted: 3 July 2004