HERMIS '92
E. Lipitakis (Editor)


# INNOVATIVE DIRECTIONS IN INFORMATION RETRIEVAL

N. N. Karanikolas and S. L. Mantzaris

Dept. of Applied Informatics
Athens University of Economics and Business
76 Patission Str., 104 34 Athens

**Abstract.** The demand for more effective information retrieval systems is expected to grow dramatically in the following years. In this paper recent trends to enhance the effectiveness of information retrieval systems are discussed and new information models are briefly described.

## 1. Introduction

Information Retrieval Systems have been used for the retrieval of Information relevant to a submitted query. The queries submitted are either in Natural Language or in a Control Query Language. In the case of a natural language query, the system automatically identifies the indexing terms embedded in the submitted query. Thus, the users of such systems does not need to be trained. On the other hand, the formulation of queries in the control query language must be done by trained experts. Independently of the way the queries are expressed, existing information retrieval systems incorporate most of the following:

- identification of the text words
- removal of stop words
- word stemming
- weighting of the word stems
- removal of the word stems with the lowest weights
- compound index term creation
- identification of associated terms and creation of thesaurus.

Salton [9] describes in detail existing models and the use of the above procedures.

In the following, recent trends in the field of information

retrieval are discussed. Section 2 discusses proposed refinements and enhancements of existing models, the central line of this section considers natural language techniques. In the last section recently proposed models are investigated.

## 2. Refinements and extensions to existing models

### Compound index term creation

Natural Language techniques are mainly used for interfacing information retrieval systems. However, there is a recent direction which uses syntactic analysis to improve the indexing ability of such systems by the use of structured terms (Waltz [14]). For example prepositional phrases, verb phrases and subject-verb patterns can be used as compound index terms. To achieve this goal the system must be able either to analyze syntactically the text sentences or to compare sequences of words against indexing patterns.

The techniques for compound term creation and usage can be divided in two main approaches. The first of them is a statistical method while the second one is a cognitive method based on the syntactic analysis of the collection documents. The first method is based on the computation of the association ratio (Church and Hanks [2]). This method views the collection of documents as a corpus. This corpus is investigated for the identification of term couples (x,y) which are in a few words distance and this fact does not seem to be accidental. The measure of the randomness is based on the association ratio of the terms which is an estimation of the mutual information (1) of the couple (x,y). In (1)

$$I(x,y)=\log_2\frac{P(x,y)}{P(x)P(y)} \qquad (1)$$

P(x) is the probability to observe x, P(y) is the probability to observe y and P(x,y) is the probability to observe the couple (x,y) in the corpus. The term couples that have mutual information greater than a given threshold are selected as compound index terms. As a next step, Church and Hanks suggested that preprocessing the corpus with a part of speech tagger or with a parser, could improve the selection of compound index terms. This preprocessing permits to investigate only certain syntactic patterns, for example subject-verb or verb-object patterns.

The second method for compound term creation is based on the assumption that a few syntactic units (for example noun phrases) are good indicators of text contents. Therefore, syntactic analysis must be used in order to determine the boundaries of the syntactic units (noun phrases) used for indexing. The problem arising is that we can have phrases with the same meaning, but with different syntactic constructions. The most desirable thing, in this case, would be a mechanism which assigns the same unique object to all different syntactic constructions that have the same meaning. This unique object could be the head of the phrase and its modifiers. Unfortunately, syntactic ambiguity does not permit us to decide if a modifier modifies the head of a phrase or is a modifier of another modifier of the head of the phrase. This can be done only with semantic interpretation (Dahlgren and McDowell [3]).

Three approaches have been tried for the solution of this problem: ignore the ambiguity, normalize the identified phrases and index by structures which incorporate the ambiguities. The first approach uses phrases directly extracted from the text. This approach does not solve but ignores the problem. The second approach automatically generates a list of indexing phrases. Then, the phrases identified in the input text are compared with the phrases of the list and classified as either exact, general (phrases in the list are constituents of the phrase identified in the input text) and novel (new phrase not found in the list). This approach normalizes the identified text phrases with phrases from the list.

The last approach uses structures that encode the structural ambiguity of the represented phrase. Matching algorithms rank the level of relevance of two such structures. There are mainly two directions to follow according to this approach. The first direction uses dependency trees (Schwartz [10]) which are trees with explicit links between each possible head modifier pair. The dependency tree for the phrase "Problems of fresh water storage and transport in containers or tanks" is presented in figure 1.

The level of match is depended on the existence of an immediate or a transitive link. For example, the phrase "storage problem" matches exactly with the dependency tree of figure 1, but the phrase "container problem" has a weaker level of match, since the word "container" has a

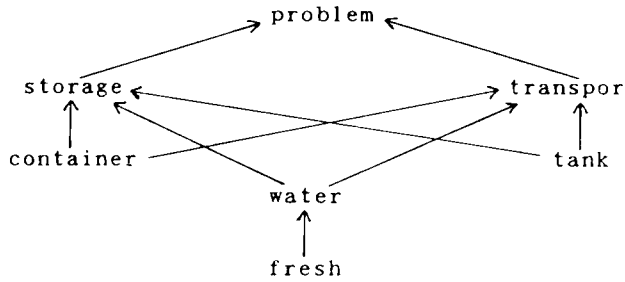


Figure 1. A dependency tree

transitive link to the word problem via the words "storage" or "transport". The second direction uses tree-structured analytics (TSAs) which are binary trees that encode structural syntactic ambiguities (Smeaton [11]). The level of match between two TSAs is depended on the words, their syntactic label, their roles in phrases as heads or modifiers and the strength of evidence for those roles by checking ambiguity markers.

Other directions for enhancing information retrieval systems

The most interesting directions for more sophisticated Information Retrieval Systems are described in the following. In the first direction, the system selected documents are presented to the user who submitted a query. Then, the user marks the most relevant of them according to his information needs. Subsequently, the system can extend the query with terms embedded in the marked documents and not present in the original query. This relevance feedback method can be repeated until the user is satisfied.

The second direction for improvements, concerns the efficiency of the retrieval strategy. Lucarella [6] suggests a nearest neighbour algorithm which minimizes the number of documents to be evaluated and also the number of inverted lists to be inspected.

The third direction is called REFORM (REad it FOR Me) and is based on the dependency structures of Schwartz [10]. Its purpose is to give

to the user of the system, a gradually more and more detailed idea about the content of a retrieved document without reading the full text.

The last direction views documents as hierarchical objects. ISO standards for document descriptions have been recently developed. The standard generalized markup language SGML (ISO standard 1986) deals with document content. It views documents as hierarchical structures. The document filling and retrieval proposal (DFR 1987) relates to the storage of document collections. It views collections as having a hierarchical structure of the type exemplified by the filling cabinet structures found in offices. As structured documents become available, as there is more need for tools which take advantage of structural knowledge. Such a tool is Maestro (Macleod [7]) which provide a query language to retrieve information contained in such structured documents.

## 3. New models

Recently, new information retrieval models have been proposed. The reasons for this proposals are :

- The need for more powerful models than the existing ones in terms of effectiveness.
- The change of the documents structure that evolve from unstructured textual objects to structured and/or multimedia objects.
- The need to embed natural language processing and artificial intelligence techniques within a coherent, controllable framework.

The importance of representing the semantics of a document for the information retrieval process is obvious. In order to refine existing information retrieval models, researchers (see e.g. Fox [4]) have attempted to use lexical-semantical relationships. Recently, Lu [5] presented an information retrieval model which is based on lexical-semantical relations. A non exhaustive list of such relations is given and these relations are grouped into five categories. The words or terms of a document that are semantically related, according to this list of relations, form hierarchical structures. A document (query) is represented as a set of these structures. To measure the structural relevance between a document and a query, a technique developed in the field of pattern recognition is used. This technique calculates the editing cost of transforming a tree T to another tree T' given three

editing operations and their costs. Experiments were carried out on very small document collection and the new model compared favorably to the well known vector space model in terms of retrieval effectiveness. However, more experiments are necessary to examine the effectiveness of this model. The most important problem with this model is that the indexing process is not yet automatic.

Recently, van Reijsbergen [8] presented a logic-based approach for information retrieval models. This approach can be regarded as a generalization of deductive databases. The underlying idea is that for a given document D to be relevant to a submitted query Q, D must logically imply Q: $D \rightarrow Q$. In other words if a query can be proved from a document, then this document is an answer to the query. In this logic-based framework, additional knowledge, which is not explicitly contained in the document, can be easily integrated in order to logically prove a query from a document.

There are two issues that must be addressed to make this approach a useful information retrieval model.

a) In order to cope with the possible existence of contradicting documents in the collection, each document is identified with a possible world W, that is a set of propositions with associated truth values.

b) Since it is impossible to compare concepts on a strict basis, van Reijsbergen introduces the notion of uncertain implication which states that D logically imply Q with certainty P: $P(D \rightarrow Q)$. Moreover, a logic for probabilistic inference is proposed in [8] which can be used to estimate the probability P, based on the amount of semantic transformation of D that is needed to prove $D \rightarrow Q$.

Another way to interpet the implication $P(D \rightarrow Q)$ was proposed by Chiaramella and Nie [1]. These two researchers presented a logic which is a combination of modal logic and fuzzy logic. In addition, they extended the initial model in [8] by considering the uncertain implication $P'(Q \rightarrow D)$ also. They considered that $P(D \rightarrow Q)$ measures to what extent the document satisfies the query and that $P'(Q \rightarrow D)$ measures to what extent the document satisfies only the query. The relevance between the document and the query can be given by combining these two measures.

The logic-based approach offers a powerful tool to investigate the

fundamental aspects of Information Retrieval. In addition, other models can be expressed in terms of this model [8]. However, much work remains to be done in developing operational systems that can use the expressive power of logic without sacrificing the efficiency.

Inference networks is another promising tool for the design of information retrieval models. Turtle and Croft [12] presented a probabilistic information retrieval model based on inference networks. An inference network is a DAG in which nodes represent propositianal variables or constants and edges represent probabilistic dependence relations between propositions. These relations are defined within the nodes. Given a set of probabilities for the source nodes (sources) of the DAG we can compute the probability of belief associated with any other of the remaining nodes.

Turtle and Croft proposed infererence networks which have two componets: a document network and a query network. The sources of the document network represent the documents. For each document there is one or more text nodes. The content of a single text node are represented by content representation nodes. Documents may share text nodes (e.g. hypertext) and text nodes may share representation nodes. The document part of the network is build once for a collection of documents. The source nodes of the query network represent the user's information need at a primitive level of concept representation nodes. There is one or more query nodes which correspond to different representations of the user's query. The query nodes are depended on concept representation nodes. All query nodes are connected to a single leaf. When a query network is build, it is attached to the document network so that document and query representation nodes match. Then, the subset of the documents which produce the highest probability for the leaf node must be selected. In general this best subset problem is intractable, but in same cases good heuristic approximations are possible.

The advantages of the inference network approach are the diversity in representing user's need with different type queries and the ability to express other models (even nonprobabilistic models like the vector space model [13]) in terms of this model. However, further work must be done towards an efficient implementation of this model to handle large scale document collections.

**References**

1) Y. Chiaramella and J. Nie, A Retrieval Model based on an Extended Modal Logic and its Application to the RIME Experimental Approach, Proc. ACM-SIGIR Conf. on Research and Development in Information Retrieval, 1990 Brussels, pp. 25 - 43.

2) K. W. Church and P. Hanks, Word Association Norms, Mutual Information and Lexicography, Computational Linguistics, Vol. 16, No. 1, 1990, pp. 22 - 29.

3) K. Dahlgren and J. McDowell, Using Commonsense Knowledge to Disambiguate Prepositional Phrase Modifiers, Proc. of the American Association for Artificial Intelligence, 1986, pp.

4) E. A. Fox, Lexical Relations: Enhancing Effectiveness of Information Retrieval Systems, SIGIR Forum, 1980, pp. 6 - 35.

5) X. Lu, Document Retrieval: A Structural Approach, Information Processing and Management, Vol. 26, No.2, 1990, pp. 209 - 218.

6) D. Lucarella, A Document Retrieval System based on Nearest Neighbour Searching, J. of Inform. Science, Vol. 14, 1988, pp. 25 - 33.

7) I. A. Macleod, Storage and Retrieval of Structured Documents, Information Processing and Manag., Vol. 26, No. 2, 1990, pp. 197 - 208.

8) C. J. van Reijsbergen, A Non-classical Logic for Information Retrieval, The Computer Journal Vol. 29, 1986, pp. 481 - 485.

9) G. Salton, Introduction to Modern Information Retrieval, 1983, McGraw-Hill.

10) C. Schwartz, Content Based Text Handling, Information Processing and Management, Vol. 26, No. 2, 1990, pp. 219 - 226.

11) A. F. Smeaton, Progress in the Application of Natural Language Processing to Information Retrieval Tasks, The Computer Journal, Vol. 35, No. 3, 1992 pp. 268 - 278.

12) H. R. Turtle and W. B. Croft, Inference Networks for Document Retrieval, Proc. ACM-SIGIR Conf. on Research and Development in Information Retrieval, 1990 Brussels, pp. 1 - 24.

13) H. R. Turtle and W. B. Croft, A Comparison of Text Retrieval Models, The Computer Journal, Vol. 35, 1992, pp. 279 - 290.

14) D. L. Waltz (Ed.), Semantic Structures Advances in Natural Language Processing, 1989, Lawrence Erlbaum Associates.