Information Dissemination based on Semantic Relations

Irini – Electra Katzagiannaki, Dimitris Plexousakis

Department of Computer Science, University of Crete
P.O. Box 2208, GR-71409, Heraklion, Greece
{ileria , dp}@csd.uoc.gr
Institute of Computer Science, Foundation for Research and Technology - Hellas
P.O. Box 1385, GR-71110, Heraklion, Greece
{ileria , dp}@ics.forth.gr

Abstract. In a selective information dissemination (SDI) system, users submit profiles consisting of a number of long-standing queries to represent their information needs. The system then continuously collects new documents from underlying information sources, filters them against the user profiles, and delivers relevant information to corresponding users. SDI systems are very important nowadays due to the vast amount of information that flows in the World Wide Web, as they inform users for relevant information, without requiring them to spend time to locate it. This paper presents an SDI system, which takes into account the lexical, as well as the semantic relationships between terms of documents and user profiles. In particular, a user profile is considered relevant to a document, if its terms or their synonyms or hyponyms appear in the document. The paper also presents a profile index structure supporting both the Boolean and Vector Space models.

1. Introduction

The majority of SDI systems [1] are based on lexical search. In particular, they represent documents and user profiles as sets of terms and check the identicalness between these two sets, in order to make the decision for delivering a document to a user. However, users of SDI systems may use many different terms to express the same meaning, terms that are called *synonyms*. Simultaneously, when users seek information about a term, they are also interested in information about terms that are *hyponyms* (specialized terms) of this initial term. As a result, it is necessary for such a system to contain mechanisms that take into account the semantic relationships between terms during matching of user profiles with documents.

This paper presents an SDI system that takes into account the semantic relationships between terms. In particular, a user profile is considered relevant with respect to a document, if its terms or the synonyms or hyponyms of them appear in the document.

The system deals with profiles that are represented in the two most popular models in information retrieval, namely the *Boolean model* ([2], [3]) and the *Vector Space model* ([4]). In order to improve the system's performance, an index structure of profiles – rather than of documents – has been created, as profile information

constitutes a larger volume and is more static. The system has been implemented and evaluated based on experiments that have been conducted, in order to validate the expected gains in providing greater precision in information dissemination.

The remainder of this paper is organized as follows: Section 2 presents the objectives of our system. Section 3 presents the methodology followed, while Section 4 presents preliminary results from the experiments carried out. Finally, Section 5 presents our conclusions.

2. Objectives

The basic goal of our system is to deliver relevant documents to users based on their interests, which are called profiles. Systems in information retrieval should have precision in delivering documents. That is documents that are sent to users must satisfy the user's needs. Our system achieves a high degree of precision by taking advantage of the semantic relations between the terms of documents and user profiles. Instead of using only the lexical match of terms, it considers the match based on semantics of terms, i.e., the possible synonymity or hyponimity relationship that may exist between terms. The information about the semantics of terms is retrieved through a thesaurus.

3. Methodology

Assuming a finite alphabet Σ, a word is a finite non – empty sequence of letters from Σ. We also assume an infinite set of words, called a vocabulary V. A term is a word from V or a combination of words from V, called a phrase. Terms have *ranks*, which denote the frequency of occurrence of each term in the document collection.

In our system, a document is represented by its metadata, that is its title, its authors and its keywords. In the Boolean model, a document consists of a collection of the keywords of the document. So, a document D with n keywords is represented as $D = (x_1, x_2, \ldots, x_n)$, where x_i, $1 \leq i \leq n$, is a term from the vocabulary V. In the Vector Space model, documents are represented as vectors of (term, weight) pairs, where term is a keyword of the document. Thus, a document D with n keywords is represented as $D = <(x_1, w_1), \ldots, (x_n, w_n)>$, where x_i, $1 \leq i \leq n$, is a term of the vocabulary V and $w_i > 0$, $1 \leq i \leq n$, is the weight assigned to the i-th term. The weight of a term denotes how statistically important it is.

User profiles consist of one or more sub-profiles, one for each topic of interest. Sub – profiles of a user profile are connected with logical OR, whereas terms are connected with logical AND. Each user profile can be viewed as a continuous query against the system. That is the profile is continuously executed and the system provides relevant documents to users as long as there are valid profiles that match these documents.

In our system, we assume that each sub – profile of Boolean model is a set of k distinct terms (y_1, y_2, …, y_k), where y_i are the terms that are connected with logical AND in a sub - profile. A sub – profile of Boolean model matches a document if each term of the sub – profile or a synonym of it or a hyponym of it appears in document.

On the other hand, sub - profiles of Vector Space model are represented as vectors of (term, weight) pairs. Thus, a sub – profile P with n terms is written as $P = <(y_1, z_1), \ldots, (y_n, z_n)>$, where y_i, $1 \leq i \leq n$, is a term from vocabulary V and $z_i > 0$ is the weight of term y_i. In Vector Space model, the degree of similarity between a document – sub - profile pair is based on the weights of the corresponding matching terms. Matching terms are the terms of the sub – profile that either they exist in document or a synonym or a hyponym of them exists in document. In order to decide if a document matches a sub - profile, the user specifies some kind of absolute relevance threshold. So for a sub - profile and relevance threshold θ, a document D is relevant if a measure, named *cosine similarity measure* [5], is greater than θ.

In a selective dissemination of information system, a vast number of profiles are stored, though documents are independently checked for matching against the profiles. As a result, for performance reasons, an index structure of profiles is required. In the index structure of our system, each term that exists in user profiles is assigned to a list of postings that consists of all or some sub – profiles that contain it. The mapping that maps a term to its list is implemented as a hash table. Sub – profiles of Boolean model appear in the list of *one* of its terms, which is called *key* [6]. Sub – profiles of Vector Space model are indexed by *some* of their terms, which are called *significant terms* [7]. Each posting in the list contains the profile identifier, the weight of the term indexed, the number of non – key terms or insignificant terms of the sub – profile respectively and the rest of the terms of the sub – profile with their weights. So, when a document arrives, it is checked against the existing index structure of profiles.

4. Experimental Results

The main goal of the preliminary experiments that were conducted was to prove that our system has improved *precision* and *recall*, which are the two most important measures that an efficient information retrieval system must satisfy.

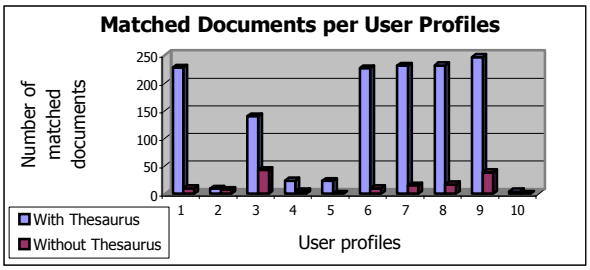

Fig. 1. Matched documents per user profile using thesaurus and not

In the experiment 1000 documents and 10 user profiles were used. The blue (lighter) bar in Fig. 1 shows the matched documents per profile in the case that matching is based on semantic relations between terms of documents and profiles (thesaurus is used). The red (darker) bar of Fig. 1 shows the matched documents per profile in the case that matching is based only on the lexical similarity between terms of documents and profiles. So, the use of semantic relations increases system recall. Other experiments, that have been conducted, have shown the improvement in precision, but due to space limitations are not presented.

5. Conclusions

The objective of the proposed system is to help users satisfy their information needs, without requiring them to spend time to locate relevant information. This push – based system receives user profiles, which are expressed using the Boolean model and the Vector Space model. While the system receives documents from underlying information sources, it filters them against the index of user profiles and delivers relevant documents to users. The benefit of the system is the improvement of the values of recall and precision, by taking advantage of the semantic relations between terms of documents and profiles. Recall is increased as a profile term is considered to be "in" a document if the term itself or a synonym or hyponym of it exists in document, while precision is increased especially in the cases that profiles contain compound terms and satisfactory number of terms.

6. References

1. T. W. Yan, and H. Garcia – Molina. Distributed selective dissemination of information. In Proceedings of the 3rd International Conference on Parallel and Distributed Information Systems (PDIS), pages 89 – 98, 1994
2. D. Gifford, R. Baldwin, S. Berlin, and J. Lucassen. An Architecture for Large Scale Information Systems. In Proceedings of the Symposium on Operating System Principles, pages 161 – 170, ACM, December, 1985
3. S. Acharya, M. Franklin, and S. Zdonik. Balancing Push and Pull for Data Broadcast. In Proceedings of the ACM SIGMOD Conference, Arizona, pages 183 – 194, May, 1997
4. D. Aksoy, and M. Franklin. Scheduling for Large – Scale On – Demand Data Broadcasting. In Proceedings of the IEEE INFO-COM Conference, San Francisco, pages 651 – 659, March, 1998
5. U. Cetintemel, M. J. Franklin, and C. L. Giles. Self-Adaptive User Profiles for Large-Scale Data Delivery. In Proceedings of the 16th International Conference on Data Engineering (ICDE), San Diego, CL, USA, pages 622 – 633, February, 2000
6. T. W. Yan and H. Garcia – Molina. Index structures for selective dissemination of information under the Boolean model. ACM Transactions on Database Systems, 19(2): 332 – 364, 1994
7. T. W. Yan and H. Garcia – Molina. Index Structures for Information Filtering Under the Vector Space Model. In Proc. International Conference on Data Engineering, pages 337 – 47, 1994