

Conceptual Matching in Web Search

using FIS-CRM for representing documents

Pablo J. Garcés
Escuela Superior Informática
Universidad de Castilla la Mancha
Ciudad Real - España
pgarces@upco.es

José A. Olivas
Escuela Superior Informática
Universidad de Castilla la Mancha
Ciudad Real - España
JoseAngel.Olivas@uclm.es

Francisco P. Romero
Soluziona Software
Ciudad Real - España
fromero@uf-isf.es

Abstract

In this paper a new approach for achieving the conceptual matching between user queries and web documents is presented. The key of the proposed system is to use FIS-CRM (*Fuzzy Interrelations and Synonymy Concept Representation Model*) to represent the indexed web pages. This model (also implemented in the FISS metasearcher) is supported by a fuzzy synonymy dictionary and various thematic fuzzy ontologies of terms.

Keywords: Conceptual Matching, Fuzzy Ontology, Fuzzy Synonymy, Web Search.

1 Introduction

Nowadays, optimizing information retrieval (IR) systems has become the main aim for most web searcher developers. Once the efficiency of search engines has been more than fully proved (to realize it, let's take a look at the time taken by Google for retrieving thousand of links), the efforts has been leaded to optimizing the quality of the results produced by searchers. It is not necessary to prove that most of the resulting links of a web query are usually not relevant to the real user aim, and worst of all, lots of documents "conceptually" related to the user query are not retrieved. To show off this later aspect let's make a query with the word "fruit". The resulting links obviously correspond to web documents that contain this word, that is, documents of "fruit growers", "fruit recipes", "botany links" and we can also find some web documents related to the results ("fruit") produced by any system. But, what about a web document containing the word "apple" or "orange"? Would this document be

retrieved by a query with the word "fruit"? Unfortunately, if the document does not contain the word "fruit" it would be never retrieved, despite containing the "concept" specified in the user query. In this sense, we can affirm that the lack of quality in the results is due to the fact that searchers only consider lexicographical aspects when matching documents and queries, but do not take into account the semantic aspects of them [9].

In order to provide a way of optimizing web search results, soft computing techniques have taken an important role [7]. In this sense, lots of approaches have been proposed in the recent last years. Some of them are leaded to the construction of flexible adaptive sites (based on web patterns, user profiles, access patterns, user behavior patterns...) using data mining techniques [4,8]. Some others are focused on organizing into groups the retrieved documents (it is important to point out those based on dynamic clustering [11] in contrast to the ones supported by predefined thematic groups).

Concerning the conceptual matching problem, different interesting approaches have been provided. A very interesting use of fuzzy techniques to information retrieval is the one proposed in [5]. The proposed system is based on the Conceptual Fuzzy Sets (CFS), which are used to represent the concepts contained in a document. Hopfield networks algorithms implement the matching mechanism. Other different group of approaches is formed by the vector space model extensions based systems. Representative of them are the ones based on WordNet, a semantic net of word groups. These systems require a special matching mechanism (ontomatching algorithm [3]) when comparing the associated concepts of the words.

The main characteristic of the conceptual matching approach proposed in this paper, on the contrary of ones mentioned before, is that it does not require a

special matching algorithm, that is, the mechanism used for matching queries and documents is the standard matching mechanism of the vector space model. The key to get this important aim is to use a model for representing web documents (FIS-CRM [6]) that is totally compatible with the standard matching mechanism of the vector space model, given that the base vector of a FIS-CRM document is merely a single term weight vector. In this sense, FIS-CRM can be considered an extension of this model that is able to represent the concepts contained in any document (instead of word occurrences). The main characteristics of this model are reviewed in section 2 of this paper.

FIS-CRM has already been implemented in other system, the FISS metasearcher [6], whose functionality is also integrated in the system proposed in this paper. In the FISS metasearcher, the model was used to represent the snippets retrieved by the search engine in order to organize the resulting links into clusters of conceptually related web pages. In section 3, the components involved in this system are presented. The system description is also accompanied with an example query.

2 FIS-CRM. Fuzzy Interrelations and Synonymy Concept Repres. Model

This model is supported by a fuzzy dictionary of synonyms (an adapted version of the one presented in [1]) and fuzzy ontologies of terms (automatically built using the algorithm presented in [10]). The fuzzy dictionary provides the fuzzy degrees of the synonymy interrelation, whereas the fuzzy ontologies provide the fuzzy degrees of the generality interrelation. FIS-CRM is based on the idea that if a word appears in a document, then its synonyms (representing the same concept) underlie it, and also the words that represent a more general concept.

The fundamental basis of FIS-CRM is to “share” the occurrences of a contained word among the fuzzy synonyms that represent the same concept and to “give” a fuzzy weight to the words that represent a more general concept that the contained one. To construct a FIS-CRM vector, first, the base weight vector (based on the occurrences of the contained words) is constructed. Afterwards, vector weights are readjusted (obtaining vectors based on concept occurrences). This way, a word may have a fuzzy weight in the new vector even if it is not contained in it, as long as the referenced concept underlies the document.

The readjustment process is implemented with an iterative algorithm that is repeated until no changes in the vector are produced. In each step, the weights of the contained words and the weights of their synonyms and more general words are readjusted (using FIS-CRM formulae [6]) taking into account the type of the involved words, polysemic or one meaning ones) and the fuzzy synonymy and generality degrees provided by the fuzzy dictionary and the fuzzy ontologies. Obviously, the context of the document plays a fundamental role when determining the meaning of a polysemic word.

3 System Components

The system has four main components that correspond with the four main steps of the global process carried out by this system. They are the web crawler (showed in figure 1), the query input component, the search engine and the clustering component.

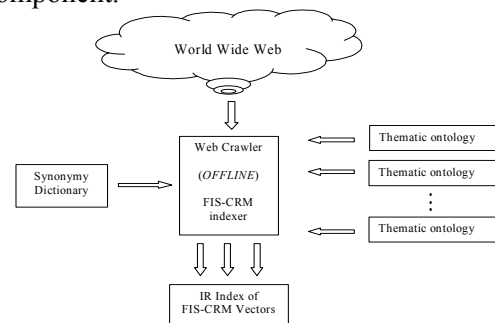


Figure 1: Web crawling offline process.

The three later components are involved in the online search process, which is showed in figure 2. Figure 3 shows the detailed data flow diagram of the global online process.

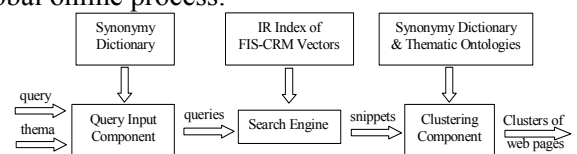


Figure 2: Query online process.

3.1 Web crawler

The process carried out by this component is offline. It includes the crawling of all the accessible web pages in the Internet and the construction of the IR index. The main characteristic of this component is that documents are indexed using the FIS-CRM model instead of using the standard vector space model. Figure 1 shows the items involved in this step. For example, a piece of an indexed FIS-CRM vector could be this one:

apple	fruit	garden	tree
2	1.4	1	1

This vector would correspond to a document containing the words “apple”, “garden” and “tree”, but not containing the word “fruit”. The fuzzy weight of this later word is obtained after applying FIS-CRM to the base vector. The piece of vector corresponding “this paper” could be this one:

web	computer	link	result	search	vector
2	1.7	1	2	5	3

The construction of the FIS-CRM vectors is carried out in the same way that the FISS metasearcher does, thus, the web crawler process consists on two clearly defined steps: the one taken by most web crawlers, that is crawling and indexing the links using the standard vector space model, and the second step that consists on applying the FIS-CRM readjusting process to the whole set of indexed documents. In this sense it is very important to emphasize that this process is carried out by an offline process not affecting the efficiency of the online search process.

3.2 Query input component

The functionality of this component and the clustering component’s one are very similar to the one carried out by the FISS metasearcher. In this case, the process carried out by the query input component corresponds to the step 1 of the global process diagram showed in figure 3. The query input component is the first component involved in the online processing of a query, and it allows the user to enter the query. The user is also required to enter the thematic ontology (if wanted) and the similarity threshold used to cluster the resulting links.

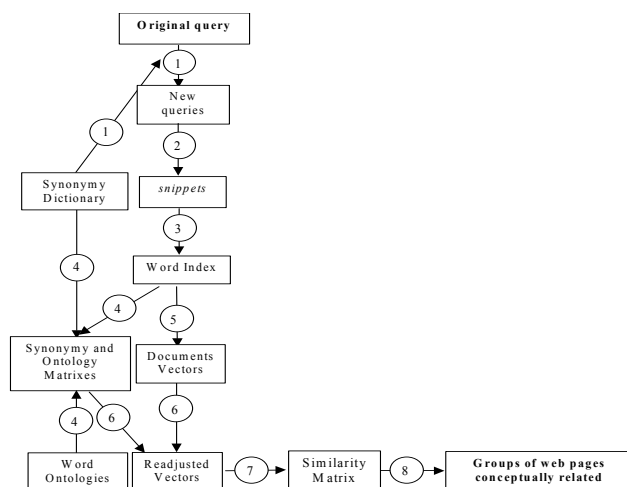


Figure 3: Detailed online process

This component generates new queries using the synonyms of the words contained in the query. Each one of the new queries has a fuzzy compatibility degree with the original query that is calculated from the synonymy degree between the words included in it and the words specified in the original query. The final value is obtained using a T-norm (product). This value is later used when ordering the links.

For example, if the user make a query with the word “fruit”, the system would generate two queries: One with the word “fruit” (with a compatibility degree of 1) and other one with the word “result”, a synonym of “fruit” (with the compatibility degree obtained from the fuzzy synonymy dictionary, in this case 0.5, which is the synonymy degree between “fruit” and “result”).

3.3 Search engine

The Search Engine is the responsible for retrieving the links of the web pages that match the generated queries. In the detailed process diagram showed in figure 3, step 2 is the one that implements the functionality of this component. Concerning the matching mechanism used by the search engine, despite the aim of the matching algorithm is getting a measure of the cooccurrences of “concepts” contained in the query and the ones contained in the indexed documents, due to the fact that the concept occurrences are spread out among the words that represent them (in the offline crawling step), *the online matching algorithm consists on a simply “word” matching process like the one implemented by most search engines*. So, the efficiency of the matching algorithm could be exactly the same that the one achieved by the most efficient search engine (if we had the same resources ...).

So, it is important to point out that, as the vectors are formed by fuzzy weights, the matching function will provide a fuzzy value for every matched web page that represents the matching degree. Just as the compatibility degree obtained previously, this value is also used when ordering the retrieved links, in the same way that the page rank obtained by the number of links to the web page is also used.

Following the before example, the corresponding documents to the vectors showed before would match the user query. As we can observe, a normal search engine would never retrieve the first one as it does not contain the word “fruit” (although the concept underlies) and the second one would not either for the same reason.

3.4 Clustering component

The functionality of this component was also implemented in the FISS metasearcher. This module implements a soft-clustering algorithm (based on the one proposed in [2]) to organize the resulting links into clusters of conceptually related web pages. FIS-CRM is also used in this component, and it is applied to the snippets retrieved by the search engine. The similarity function contemplates the co-occurrence of words and phrases, size of the snippets, rarity of words and the co-occurrence of concepts (in an implicit way). In this step, groups are also labelled with the words that represent the concepts inside the group's documents. Steps 3 to 8 of the data flow diagram showed in figure 3 describe the detailed process carried out by this component. The documents retrieved by the example query would be organized in various overlapped clusters (one related to "botany", other one related to "gastronomy", etc).

4 Conclusions

The main aspect of the FIS-CRM model is that it may be easily integrated in any searcher since this model provides an extension of the vector space model that is totally compatible with the standard matching algorithms used in most search engines.

At present, FIS-CRM efficiency has been successfully proved in the FISS metasearcher [6], and it will be shortly proved when implemented in the system proposed in this paper, and really we are very confident on getting the desired aim.

Our trust is based on: the results produced by the FISS metasearcher when making conceptual clusters, the simplicity of implementing FIS-CRM in the web crawler component, and it is also based on the fact that the matching mechanism is exactly the same that most search engine have, not having to get its efficiency decreased, and for these reasons, we sincerely think that the proposed approach provides a very useful contribution to the semantic web search field that any search system could easily integrate.

Acknowledgements

This research is supported by CIPRESES (TIC 2000-1362-C02-02) Project, MCYT, Spain.

References

- [1] S. Fernandez, A contribution to the automatic processing of the synonymy using Prolog, *PhD Thesis*, University of Santiago de Compostela, Spain, 2001.
- [2] L. King-ip, K. Ravikumar, A similarity-based soft clustering algorithm for documents, *Proc. of the Seventh Int. Conf. on Database Sys. for Advanced Applications*, (2001).
- [3] A.K. Kiryakov, K.I. Simov, Ontologically supported semantic matching, *Proceedings of "NODALIDA'99: Nordic Conference on Computational Linguistics"*, Trondheim, (1999).
- [4] M.J. Martin-Bautista, M. Vila, D. Kraft, J. Chen, User profiles and fuzzy logic in web retrieval, *Proc. of the BISC Int. Workshop on Fuzzy Logic and the Internet*, (2001), 19-24.
- [5] R. Ohgaya, T. Takagi, K. Fukano, K. Taniguchi, Conceptual fuzzy sets- based navigation system for Yahoo!, *Proc. of the 2002 NAFIPS annual meeting*, (2002) 274-279.
- [6] J.A. Olivas, P.J. Garcés, F.P. Romero, FISS: Application of fuzzy technologies to an Internet metasearcher, *Proceedings of the 2002 NAFIPS annual meeting*, (2002) 140-145.
- [7] G. Pasi, Flexible information retrieval: some research trends, *Mathware and Soft Computing* 9, (2002) 107-121.
- [8] M. Perkovitz, O. Etzioni, Towards adaptive web sites: Conceptual framework and case study, *Artificial Intelligence* 118, (2000) 245-275.
- [9] I. Ricarte, F. Gomide, A reference model for intelligent information search, *Proc. of the BISC Int. Workshop on Fuzzy Logic and the Internet*, (2001) 80-85.
- [10] D. Widiantoro, J. Yen, Incorporating fuzzy ontology of term relations in a search engine, *Proceedings of the BISC Int. Workshop on Fuzzy Logic and the Internet*, (2001), 155-160.
- [11] O. Zamir, O. Etzioni, Grouper: A dynamic clustering interface to web search results, *Proceedings of the WWW8*, (1999).