Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

ELSEVIER International Journal of Approximate Reasoning 34 (2003) 201–219

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

www.elsevier.com/locate/ijar

An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents

José A. Olivas [a,*], Pablo J. Garcés [a], Francisco P. Romero [b]

[a] *Department of Computer Science, UCLM, Paseo de la Universidad 4, 13071–Ciudad Real, Spain*
[b] *Soluziona, Centro Mixto de Investigación y Desarollo Soluziona–UCLM, Ronda de Toledo s/n, 13003–Ciudad Real, Spain*

Received 1 March 2003; accepted 1 July 2003

Abstract

The main objective of this work is to improve the quality of the results produced by the Internet search engines. In order to achieve it, the FIS-CRM model (Fuzzy Interrelations and Synonymy based Concept Representation Model) is proposed as a mechanism for representing the concepts (not only terms) contained in any kind of document. This model, based on the vector space model, incorporates a fuzzy readjustment process of the term weights of each document. The readjustment lies on the study of two types of fuzzy interrelations between terms: the fuzzy synonymy interrelation and the fuzzy generality interrelations (''broader than'' and ''narrower than'' interrelations). The model has been implemented in the FISS metasearcher (Fuzzy Interrelations and Synonymy based Searcher) that, using a soft-clustering algorithm (based on the SISC algorithm), dynamically produces a hierarchical structure of groups of ''conceptually related'' documents (snippets of web pages, in this case).

Keywords: Internet search; Fuzzy synonymy; Fuzzy generality; Concept representation model; Metasearcher; Soft-clustering algorithm

* Corresponding author.
E-mail addresses: joseangel.olivas@uclm.es (J.A. Olivas), fpromero@soluziona.com (F.P. Romero).

0888-613X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2003.07.008

1. Introduction

Most users of the Internet searchers often complain about the lack of quality of the obtained results, and we are all used to have to filter the retrieved links in order to identify the ones really related to the aim of our query. In this sense, we can affirm that most of the resulting pages have no relevance to the user.

For example, if you try to find the related pages to "matrix", you should get links about computers, medicine, anatomy, the film "Matrix"..., but none about arrays or vectors (assuming that computer science was the aim of our search).

In a similar way, if you make a search with the word "*sport*" you will retrieve pages containing this word, but you will not retrieve some others that, despite not containing this word, do contain other words ontologically related to it, such as "*basket*" or "*football*".

In both cases, the lack of quality in the results is due to the fact that searchers only consider lexicographical aspects when comparing the words of the query with the words contained in the web documents, but do not take into account the semantic aspects of them [1]. In this sense, this consideration would clearly contribute to improve the development of the concept of *Semantic Web* [2].

In order to provide a solution to this problem, lots of approaches have been proposed in the last years, and many of them have been gathered in these very interesting papers [3–5].

These approaches try to deal with the problem from different points of view, and vary from the natural language processing trends to flexible and adaptive systems approaches (most of them based on soft computing techniques), passing through the vector space model based information retrieval (IR) systems and the ones based on fuzzy logic capabilities. In fact, we can find some works that could be catalogued in various of these trends.

The analysis of the natural language [6] is an important field that, in fact, underlies all of the IR approaches and, unfortunately, still constitutes a challenge to science in general and to AI in particular [7].

Among the ones based on the vector space model [8], it is important to point up the systems that are supported by a thesaurus, like Wordnet [9], that provides a semantic net of concepts based on the synonymy and the hypernymy interrelationships. An example of this kind of system is the one proposed in [10] that uses an index of three numbers to attach to each word of a document its right lexical concept (called synset in Wordnet). In [11], the onto-matching algorithm is proposed for matching the words in the query with the words in the document Obviously, the main handicap of these techniques is to disambiguate the words with several meanings [12], using for this purpose a collection of hand tagged documents that provide for each word its right synset. In this sense, in [13] is proposed a corpus based algorithm to disambiguate words,

that considers the probability that certain concepts occur together. Related to the vector space model but in a different direction of the ones mentioned before, we can find an IR system [14], that treats the relative synonymy (synonymy between two words respect to a third word) as the basis for the construction of conceptual vectors.

Concerning to the construction of flexible and adaptive sites, data-mining techniques and soft computing methodologies play a fundamental role when building patterns of users, behaviours or web pages. For example, the SCML clustering algorithm and the Pagegather algorithm [15,16] allow personalized indexes of web pages to be synthesized from the visitor access patterns (obtained from user access logs). In [17] a system that builds user patterns using data-mining techniques is proposed; in this case, the patterns are used to adapt the queries according to the user profile. Another important aspect that has focused many approaches is the effort in organizing the results produced by the searchers in a suitable way for the users' expectations. The STC clustering algorithm [18] allows the resulting links of a query to be grouped into clusters obtained from the co-occurrence of words contained in the snippets retrieved by the search engine. In [19] neural networks (Kohonen maps) have been used to cluster documents and to retrieve the documents similar to one specified. Genetic algorithms have been mainly applied to IR for improving document representation and indexing [20,21] and relevance feedback [22].

Although all of these approaches provide an important improvement for the results of the IR systems, and as Prof. Zadeh [23] wrote to the the members of the BISC group, in Internet almost everything is approximate in nature, so modes of reasoning should be approximate rather than exact and, by this means, in order to achieve a semantic web, the approximate reasoning capability of Fuzzy Logic is the fundamental key.

The way that Fuzzy Logic can be applied to IR systems vary quite much from one systems to others. It can be used for defining flexible query languages [5,24–26] that allow soft constraints to be defined in the query. It is also highly useful when evaluating and filtering on the web [27]. In [28] a fuzzy generalization of the boolean model is presented, allowing aggregation operators as linguistic identifiers. Considering partial matching [29] between the terms in the query and the terms in the document is other frequently use of the fuzzy logic. It is also used in soft-clustering algorithms such as the SISC algorithm [30]. In [31] user profiles are fuzzy represented in order to expand user queries. For a similar purpose, fuzzy ontologies of terms [32] or fuzzy pseudo-thesauri [33,34] can be used. Conceptual Fuzzy Sets (based on Hopfield networks) are proposed to achieve conceptual matching dealing with context-dependent word ambiguity [35] and implemented in navigation system for yahoo! [36].

Our work might be included in all the trends mentioned before because, as it will be described later on, has to manage with natural language processing, it is based on the vector space model, it uses soft computing methodologies (soft

clustering algorithms) and, of course, is supported by fuzzy logic for representing and clustering the web pages.

This work has been focused on two main topics: The definition of a model (FIS-CRM) for representing the concepts contained in any kind of document and the construction of a metasearcher (FISS) capable of making clusters of conceptually related web pages.

In Section 2 FIS-CRM is widely described, making an special effort in explaining the readjustment process of the term weights of the document vectors. Its aim is that the weight vectors represent occurrences of concepts instead of occurrences of words. This fuzzy process is based on the study of the fuzzy synonymy relationship and the fuzzy generality relationship.

In Section 3 all the processes involved in the resolution of a user query by the FISS metasearcher are described in detail. Although the term "metasearcher" is normally used when referring to search tools that use several search engines to retrieve web links from different sources and then fuse them in an only list, the proposed system, FISS, today integrates only one search engine. Therefore, the term "metasearcher" is used simply to define a search tool that works above a search engine.

Section 4 contains some examples to show the efficiency of the results obtained with the model and the metasearcher. The best proof of their possibilities is the fact that two documents with no relevant words in common are considered similar (and included in the same cluster) because of the common concepts contained in them.

Finally in Section 5, conclusions and future trends are described. As it will be described later on, the most important future improvement is that, at indexing time, the web crawler apply FIS-CRM to the whole stored web pages (instead of applying it only to the retrieved ones) in order to make possible to any search engine to retrieve the web pages that are conceptually related with the user query.

2. Fuzzy interrelations and synonymy based concept representation model

Standard mechanisms of representing documents are usually based on considering the frequency of the contained words. Thus, the comparison of documents is restricted to the co-occurrence of words, not taking into account the meaning of the involved words.

In order to extract a part of the semantic richness of a document, it is highly useful to study the interrelations among the contained words. Obviously, this process is hindered by the natural language ambiguity, which is increased by the huge variety of language resources, such as polysemy, metaphors, etc.... Therefore, to mitigate these ambiguities it is essential to study the context of the documents.

FIS-CRM, the representation model we propose, is based on two main points:

(a) If a word appears in a document, its synonyms that represent the same concept underly it.
(b) If a word appears in a document, the words that represent a more general concept underly it.

Although actual soft-computing trends seem to question the usefulness of the vector space model (VSM) for representing documents, FIS-CRM is based on this model. In contrast to other models (VSM based) which require a complex extension of the document vectors to represent the contained concepts (and consequently an special matching mechanism) [11], the base vector of a FIS-CRM document is merely a single term weigth vector that is totally compatible with the standard matching mechanisms of most searchers.

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 obtain this aim, documents must be first represented by their base weight vectors (based on the occurrences of the contained words) and afterwards, a weight readjustment process is made to obtain a new vector (based on concept occurrences). In 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.

To carry out the readjustment, the synonymy and generality fuzzy interrelations has to be taken into account.

2.1. Synonymy fuzzy interrelation

Recent studies in the field of synonymy (gathered in [37]) have shown that the synonymy relation has a fuzzy behaviour in the sense that it is a gradual relation. So, different measures (jaccard coefficient, cosine coefficient...) can be used to establish the certainty degree of the synonymy interrelation between two words.

For example, the synonymy degree (SD) between two words A and B could be calculated using the expression below.

$$\mathrm{SD}_1(A,B) = \frac{\text{Number of synonyms held in common by } A \text{ and } B}{\text{Number of total synonyms of } A \text{ and } B} \tag{1}$$

In this case, the degree represents how synonym is B respect to A, and $\mathrm{SD}(B,A)$ represents how synonym is A respect to B.

The fuzzy synonymy degrees between every pair of synonyms (in Spanish language) obtained with this expression are available in a fuzzy synonymy dictionary [37].

As it can be observed, the synonymy relationship obtained with this expression is symmetrical, and do not consider the number of meanings of the involved words. So, in this work, in order to take account of it and to get a non symmetrical relationship (assuming that it is the real behaviour of this relationship) the expression has been adapted to this one:

$$\mathrm{SD}_2(A,B) = \frac{\mathrm{SD}_1(A,B)}{\text{Number of meanings of } A} \tag{2}$$

For example, assuming that word A has two meanings, with this formula, the measure of how synonym is B respect to A, $\mathrm{SD}(A,B)$, will take into consideration the probability (0.5 in this case) of word B to be a synonym of word A in the context of the document (as the word A could refer to a different meaning than B's one).

On the other hand, the synonymy relationship is not transitive, unless strong words were involved. Only in this case, if C is synonym of B and B is synonym of A, we can affirm that C is synonym of A.

The vector readjustment made using the synonymy interrelation is hindered by the fact that there are lots of polysemic words (words with several meanings). In this work these words are called "weak" words, whereas words with just one meaning are called "strong" words.

Depending on the types of words involved we can distinguish two types of readjustment: Readjustment between strong words and document context dependent readjustment.

2.1.1. Readjustment between strong words

It is applied whenever a strong word appears in a document and a synonym of this, which is also strong, appears in other document of the retrieved collection of documents.

For example, assuming that "car" appears twice in the document $D1$ and "automobile" appears three times in the document $D2$ (both words are strong synonyms), the corresponding base vectors (ignoring all the others words contained in them and before making any readjustment) would be these ones:

	Car	Automobile
D1	**2**	**0**

	Car	Automobile
D2	**0**	**3**

Now, if we we use a typical function (like the one showed in Fig. 3) to get a measure of the similarity of both documents, there would not be any similarity at all between these documents (as $2\times0+0\times3$ gives 0 co-occurrences), although they contain two words that represent the same concept.

$$\text{Similarity}(D1, D2) = \sum_{i}^{N} \text{weight}(t_i, D1) * \text{weight}(t_i, D2) \tag{3}$$

where

- t_i: Each one of the terms in the document N-vectors.

Assuming that both words have only one meaning and they have not other synonyms, the synonymy degree between them is 1, and the readjustment is made using the expression below:

$$\text{Weight}(x) = C(x)\sqrt{\frac{1}{T(x)}} \tag{4}$$

where

- $C(x)$ is the number of occurrences of all the words that represent the same concept that the word 'x'.
- $T(x)$ is the number of words that represent the same concept than 'x'.

Using this expression, the occurrence of the concept (vehicle moved on wheels...) is shared among the words that represent that concept. The expression uses the square root as a way of getting the expected co-occurrences when comparing the documents with an expression (Fig. 3) that is based on a sum of products. So, after making the readjustment the vectors would be like these:

	Car	Automobile
D1	**1.414**	**1.414**

	Car	Automobile
D2	**2.121**	**2.121**

With these vectors, when comparing the documents (with the similarity function of Fig. 3), you will get 6 co-occurrences (obtained by $1.414 \times 2.121 + 1.141 \times 2.121$) as we wished (2 occurrences of the concept in $D1$ and 3 occurrences of the concept in $D2$ must produce 6 co-occurrences).

2.1.2. Document context dependent readjustment

This kind of readjustment is applied whenever a weak word and a strong synonym of it appear in a document. To illustrate what happens when a word ("table" in this case) has several meanings (weak word), let us consider this vector:

Desk	Table	Chart
0	**1**	**0**

The corresponding document of this vector contains the word "table" but, contrary to what we have made in the before example, we can not share the

occurrences of the word "table" among its strong synonyms ("desk" and "chart") because its meaning in this document is not clearly defined.

Other case would be if the document also contained the word "desk", because we could consider (assuming some risk) that the word "table" would refer to the same meaning as "desk".

The vector (before readjustment) would be this one:

Desk	Table	Chart
1	**1**	**0**

Now, we share the occurrences of "table" with its strong synonym "desk" that refers to their common meaning. In this case, when sharing the concept occurrences, the weight of the strong word (desk) should increase and the weight of the weak word (table) should decrease, in order to increase the similarity of this document with other documents containing the concept associated to the word "desk".

The readjustment is made using these expressions:

$$\begin{aligned} \text{Weight}(s_i) &= C(i) + C(w)\sqrt{\frac{\text{SD}(w,s_i)}{1+\text{SDSum}}} \\ \text{Weight}(w) &= C(w)\sqrt{\frac{1}{1+\text{SDSum}}} \end{aligned} \tag{5}$$

where

- $C(w)$ is the number of occurrences of the weak word w.
- $C(s_i)$ is the number of occurrences of the strong word s (synonym of w), that is also present in the document.
- $\text{SDSum} = \text{SD}(s_i, w)$ for all the strong synonyms (s_i) of w that appear in the document.

We can observe that, with these formulae, the weight of the weak word 'w' ("table" in the example) gets decreased in favour of the strong words s_i ("desk" in this case) contained in the document. The devaluation is obtained as the new weight of the weak word is inversely proportional to the sum of the similarity degree of the strong words (that will take a "piece" of this weight), which is totally consistent to the fact that each strong word increases its weight proportionally to its synonymy degree with the weak word. So, assuming that $\text{SD}(\text{desk}, \text{table}) = 0.5$ and $\text{SD}(\text{table}, \text{desk}) = 1$, we get this new vector:

Desk	Table	Chart
1.577	**0.816**	**0**

With this vector, the document will increase its similarity with other documents that contain the word "desk", decreasing its similarity with the ones

that, despite containing the word "table", had not their meaning clearly defined.

2.2. Fuzzy ontological interrelation

Besides synonymy, there are other types of interrelations that allow us to improve the knowledge about the semantic content of a document. Although considering as much fuzzy interrelations as possible could be useful in order to exploit the semantic component of a document, in the system we present in this paper, the ones that have been considered are the generality interrelations: "broader than" and "narrower than", which are usually implemented in ontologies or hierarchical classifications of words. So, in this paper an ontology may be considered as a set of related trees where each node represents a word. In these trees, node A is descendant of node B if the word A is "more concrete than" or "narrower than" the word B, or if the word B is "more general than" or "broader than" the word A.

Obviously, the generality relationship is not symmetrical, but it can be considered to satisfy the transitivity property when the chain of words belong to the same ontology.

Widyantoro and Yen [32] propose an algorithm to automatically obtain ontologies from a collection of selected documents about a concrete field (the construction of a generic ontology is widely refused). This algorithm also allows us to get a fuzzy measure of the generality degree between each pair of words contained in the document collection, using this expression:

$$GD(A,B) = \frac{\text{Number of co-occurrences of } A \text{ and } B}{\text{Number of occurrences of } B} \tag{6}$$

In this work, the fuzzy interrelation of generality is used by the FISS tool for a completely different purpose to the one given by Widyantoro and Yen.

Other interesting approaches on fuzzy associations and their application to information retrieval are the ones proposed by Miyamoto [33,34]. In this case, the fuzzy associations are kept in fuzzy pseudo-thesauri.

2.2.1. Vector readjustment due to the generality interrelation

This type of readjustment makes possible that a word A, not contained in a document, gets a weight if a word B (narrower than A) is present in the document. In this case, the readjustment made is linear and proportional to the generality degree between A and B.

For example, the vector of a document that contains the strong word "football" but not the word "sport" would be like this:

Football	Sport
2	**0**

In this way, this document would not be retrieved by a query with the word "sport". Assuming that the word "football" is "narrower than" the word "sport" with a certainty degree of 0.4 (i.e.), the vector would be readjusted this way:

Football	Sport
2	**0.8**

By this means, this document could be retrieved using a query with the word "sport" and, what is used for in this work, it will get its similarity with other "sport documents" increased.

As it occurs with synonymy, to make this kind of readjustment we must take into account the word types, being necessary to consider the context when weak words are involved. In this sense, assuming that the different meanings of a word belong to different fields, and assuming the availability of a term ontology for each one of this fields, a weak word is readjusted using the ontology that is associated to the document. Identifying the ontology associated to a document may be a simple task of comparison the words in the document with the words in each ontology, and could be made at indexing time.

2.3. Weights readjustment algorithm

The readjustment process of the base vector weights is carried out by an iterative process that is repeated until no changes in the vector are produced. So, for each one of the contained words in the document, the corresponding formula (depending of the kind of word) is applied in order to get its new weight and the weight of its synonyms and more general words.

This process is supported by an auxiliary matrix that avoid to repeat a readjustment between the same pair of words. The restrictions defined on the two types of synonymy readjustment avoid undesired transitive synonyms to be applied (as the formulae are only applied when the involved words are strong or when the document context lets identify the right meaning), while transitive chains of applications of the generality readjustment are applied until the top word node of the tree is reached. That is, if the word A is contained in the document and B is more general word than A, B will get a weight (proportional to their generality degree), and if B has a more general word, C, this latest one will get a weight too, and so on until the most general word.

3. FISS: Fuzzy Interrelations and Synonymy based Searcher

FISS has been the first application of the FIS-CRM model and is characterized by the following aspects (Fig. 1):

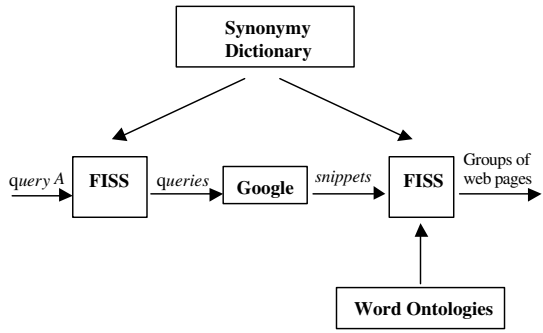

Fig. 1. Global search process.

- It generates alternative queries (from the original query) using a fuzzy synonymy dictionary.
- It uses the services of other searcher (Google in this case) to look for the links to web pages.
- It constructs a conceptual vision of the documents using for this purpose: the snippets returned by Google, the fuzzy synonymy dictionary and the fuzzy ontologies and, of course, using FIS-CRM for representing the documents.
- It dynamically groups the retrieved documents using a soft clustering algorithm that considers the co-occurrence of concepts, labelling the resulting clusters with their more relevant concepts.

The detailed search process showed in Fig. 2 is divided into the following steps:

1. *Generation of new queries*. Alternative queries are generated from the synonyms of the words in the query. Each one of these queries has the same number of terms than the original query, and is constructed from the combination of the original terms and their synonyms (keeping the boolean operators specified in the original 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 (obviously the original one is also considered and it has a compatibility degree of 1). The value is obtained using a T-norm (product in this case).
2. *Processing queries and obtaining snippets*. The generated queries are sent to Google and afterwards the snippets of the documents are extracted from the returned HTML code.
3. *Constructing an index* of words that includes all the words included in the snippets. At this step, words are pre-processed, excluding the ones in a stop

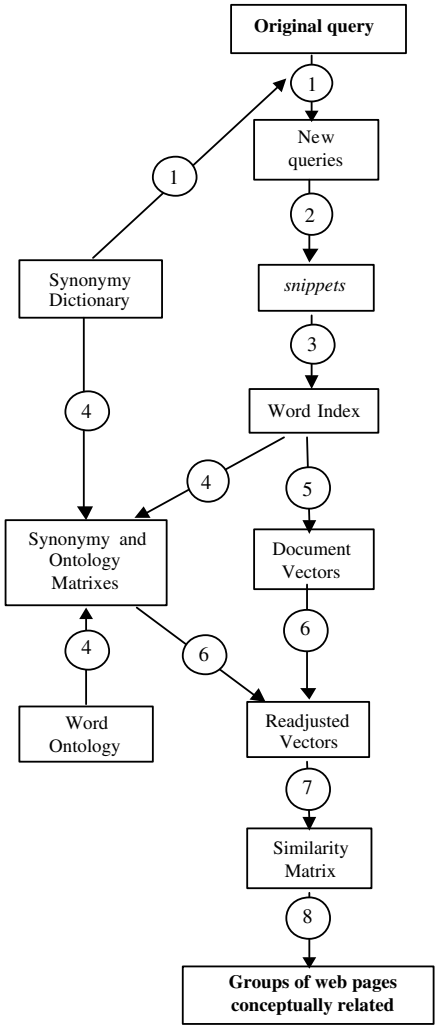

Fig. 2. Detailed search process.

list, converting them into lower case form, changing them into to singular form, considering the phrases put between quotes as individual items, etc.. . .

4. *Constructing the synonymy and the ontology matrixes*, which store the synonymy degree and the generality degree between every pair of words in the index.
5. *Representing the snippets* as vectors. The weight of each term is proportional to the number of occurrences.
6. *Readjustment of the vectors weights* using FIS-CRM.

7. *Generating the similarity matrix* that stores the similarity degree between every pair of documents in the collection.

$$\text{Similarity}(D1, D2) = \frac{\sum_i^N \text{weight}(t_i, D1) * \text{weight}(t_i, D2) * \text{rarely}(t_i)}{\text{size}(D1) * \text{size}(D2)} * \text{PositionFactor}$$

The similarity function contemplates the co-occurrence of words and phrases, size of the snippets, rarely of words and the co-occurrence of concepts (in an implicit way). If the *PositionFactor* would not have been included, two documents with the same words, but in different positions, should take a similarity of 1. So, the *PositionFactor* is a value that can only reach the value of 1 when the words in both documents have the same position.Taking the stored positions of the words in the documents from the index, it is possible to match phrases in both documents, and so calculating this factor.

8. *Generation of groups of web pages conceptually related.* In this step, groups are also labelled with the words that represent the concepts inside the groups documents.

FISS metasearcher implements a clustering algorithm that, although it is based on the SISC algorithm [30], has been adapted to optimize the resulting groups and to be able to get a hierarchical structure of groups. This algorithm is characterized by creating an initial number (automatically calculated) of centroid clusters, followed by an iterative process that includes each document in the clusters whose average similarity is upper than the threshold of similarity (automatically calculated, but user specified if wanted). The threshold is calculated from the average similarity between each document to all the clusters and its maximum similarity to a cluster. The algorithm also consider merging clusters and removing documents for clusters when their average similarity decreases under the threshold. In order to get a hierarchical structure, big clusters and the bag cluster (formed by the less similar documents) are reprocessed with the same method. Concerning the characteristics of the obtained results:

- The whole collection of documents retrieved can be considered as a fuzzy set that is formed by the documents that would have been obtained using the original query and the ones obtained from the generated queries. The first ones' membership degree to the fuzzy set is 1, and the second ones have a membership degree that is equal to the compatibility degree of the query that produced them with the original query.
- The resulting clusters can be considered as fuzzy sets, so each one of the documents retrieved has a membership degree (obtained from an average

similarity degree) to each one of these clusters. To make the solution easy to understand, the documents that have a membership degree to a cluster lower than the similarity threshold are not included in this cluster.

Ranking criterion. When presenting the links of each cluster to the user, the membership degree of each snippet to the cluster can be used as a ranking criterion. In this sense, it is important to remark that the ranking information provided by the search engine could also be combined with this value.

4. Tests and validation

FISS actual version uses a Spanish synonyms dictionary and a Spanish gastronomic ontology. The dictionary was obtained from the one provided by Dr. Fernández [37] and the ontology was constructed from a collection of web pages about gastronomy using the method proposed in [32].

The tests made with the metasearcher FISS have produced interesting results as we will show using the example of a query with the word "judías" (Spanish word with two meanings: "beans" and "Jewish"). Fig. 3 shows the snippet of one of the retrieved links (a recipe of "beans soup with bacon"):

Fig. 4 allows to compare the vector weights before and after applying FIS-CRM. We can see that some words, like "hortaliza" (vegetable), have got a weight (that they had not before) because of their ontological relation with the word "berza" (cabbage). On the other hand, "panceta" (bacon) has taken a weight at the expense of "tocino", whose weight is decreased ("panceta" and "tocino" are synonyms of bacon). The same situation occurs between the words "berza" and "col", both Spanish synonyms of "cabbage".

Fig. 5 shows the semantic expansion produced when readjusting the vector (the area represents the not contained words that have got a weight).

The best way to test the consequences of applying FIS-CRM is to calculate the similarity between two documents (with the same similarity function) in two ways: The first one using the base vectors of the documents and the other one using the readjusted vectors. The result of this comparison is that, in some cases, the similarity increases using the readjusted vectors, but other times the similarity decreases. The first case is due to the fact that the semantic expansion suffered by the documents has come one close to the other, and in the second

Crema de **garbanzos** o de **alubias**
.. **Garbanzos** o **lubias Costilla** y **tocino** de **cerdo** Huesos de...
garbanzos con sal, la **costilla**, el **tocino** y la **berza**. Sacamos la **costilla** y el **tocino** y trituramos el ...
teleline.terra.es/personal/aiolozil/r30.htm - 8k -En caché - Páginas similares

Fig. 3. Snippet of a retrieved document.

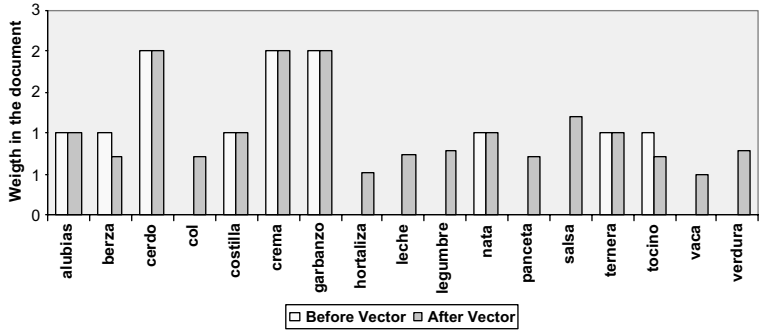

Fig. 4. Vector weights before and after the readjustment.

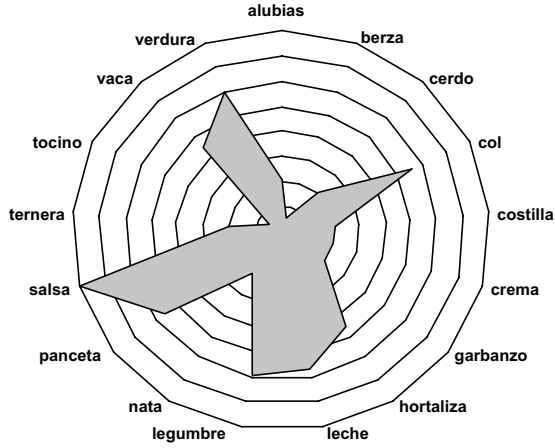

Fig. 5. Semantic expansion of a document.

case, the weights depreciation occurred when weak synonyms are involved (supported by the context) makes the meaning of the documents being farther. Fig. 6 shows one of the groups of links obtained at the end of the process.

This group contains eight conceptually related documents, although only three of them are shown in the main page. They all refer to Jewish people, as it is one of the meanings of the word "judías". At the top of Fig. 6 we can see the words used to label the group. The number in brackets that follows each word means the number of documents in the group that contain the concept that this word represents (but the word itself may not be present).

This group also allows us to test that one document may belong to more than one group, as it happens with the one titled "Indice de recetas judías"

Grupo Nº0 Cantidad de páginas:8

israeli(8) hebreo(8) judia(7) fiestas(3) gastronomia(3)

Indice de Recetas Judías
Click here! Discount Travel Hebreos Net -Cocina Hebreos Net, Recetas de la cocina judía. Beiguelej; ...
www.hebreos.net/Cocina/

Nota aparecida en el Diario "El País"- España
... Cartas Judias www.hebreos.net correo@hebreos.net #hebreos (www.undernet.org) Unir a los judíos hispanohablantes del mundo es el objetivo de esta lista de ...
www.hebreos.net/Info/elpais0700.htm

seder
... Fiestas Judías. 1- El "Seder" (orden), la "Cena Pascual", es la gran comida de Rey que celebran los judíos el día primero de la Pascua, donde Jesús ...
biblia.com/catolicos/seder.html

Mostrar Todas las Páginas del Grupo 0

Fig. 6. Group produced by FISS.

(catalogue of Jewish recipes) which is also included in a group (not shown) of recipes with beans.

To show the effectiveness of the semantic expansion produced by the re-adjustment let us consider the pair of snippets below (Fig. 7).

The first snippet is about a recipe of ''Alubias con chorizo'' (beans with a Spanish exquisite kind of pork sausage) and the second one is about a recipe of ''Pochas con panceta'' (beans with bacon). Both documents do not have in common any relevant word, and due to this, a basic similarity function would not produce any similarity at all. But if we apply FIS-CRM to the documents (readjusting their weights), they get a similarity degree of 0.3. This is due to the synonymy interrelation among ''judías'', ''alubias'' and ''pochas'' and above all, because of the ontologic interrelation between ''panceta'' and ''chorizo''

Chef Uri : Entremetier. Alubias con Chorizo
... Alubias con Chorizo. sal; pimienta Poner las alubias en remojo la noche anterior. Ponerlas en una olla cubiertas con agua y cocerlas junto con el chorizo. En ...
www.chefuri.com/java/entre/alubiasconchorizo.htm

Pochas con panceta. Recetas. Especial Gastronomía. EL CORREO ...
... EL MENÚ / RECETAS Pochas con panceta. Ingredientes para dos raciones: 300 g. de pochas; 150 g. de panceta de cerdo ibérico ahumada; 1 tomate; 1 pimiento verde ...
www.diario-elcorreo.es/gastronomia/recetas/receta290700.html

Fig. 7. Two conceptually related snippets.

with "cerdo" (pork), as we consider the words "panceta" and "chorizo" narrower than the word "cerdo".

5. Conclusions and future work

Despite the identification and representation of concepts being nowadays an utopia, the model introduced in this work (FIS-CRM) allows us to get, at least, a part of the potential semantic richness that underlies in every document.

The main aspect of this 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.

Concerning the model itself, at this moment we are trying to improve some of its aspects such as managing other type of interrelations and studying more mechanisms to determine and exploit the context (when managing weak synonyms).

With regards to the metasearcher, there are also lots of aspects capable of improvement, such as considering more languages, using more ontologies (the user could choose one when formulating the query), improving the pre-process of words when indexing (reducing the words to their lexical root), optimizing the storage requirements and some algorithms and considering the fuzzy deformable prototypes [38] for clustering processes.

Anyway, the most important aspect to be considered is to apply FIS-CRM to all the accessible pages (the ones indexed by the web crawler), instead of applying it to the retrieved ones.

Thus, any search engine that integrates the FIS-CRM readjustment mechanism when indexing, will be able to retrieve the pages "conceptually related" to the ones included in the query and what is more important, not having to modify the search mechanism and not decreasing the efficiency of the search process.

References

[1] I. Ricarte, F. Gomide, A reference model for intelligent information search, in: Proceedings of the BISC International Workshop on Fuzzy Logic and the Internet, 2001, pp. 80–85.
[2] D. Fensel, M. Munsen, Special issue on semantic web, IEEE Intelligent Systems (IEEE IS) 16 (2) (2001).
[3] M. Nikravesh, Fuzzy conceptual-based search engine using conceptual semantic indexing, in: Proceedings of the 2002 NAFIPS Annual Meeting, 2002, pp. 146–151.
[4] M. Kobayashi, K. Takeda, Information retrieval on the web, ACM Computing Surveys 32 (2) (2000) 144–173.
[5] G. Pasi, Flexible information retrieval: some research trends, Mathware and Soft Computing 9 (2002) 107–121.

[6] A. Smeaton, Progress in the application of natural language processing to information retrieval tasks, The Computer Journal 35 (3) (1992) 268–278.
[7] T.H. Cao, Fuzzy conceptual graphs for the semantic web, Proceedings of the BISC International, in: Workshop on Fuzzy Logic and the Internet, 2001, pp. 74–79.
[8] G. Salton, A. Wang, C.S.A. Yang, Vector space model for automatic indexing, Communications of the ACM 18 (1975) 613–620.
[9] G.A. Miller, WordNet: A lexical database for English, Communications of the ACM 11 (1995) 39–41.
[10] J. Gonzalo, F. Verdejo, I. Chugur, J. Cigarran, Indexing with WordNet synsets can improve retrieval, in: Proceedings of the COLING/ACL Work. on usage of WordNet in natural language processing systems, 1998.
[11] A.K. Kiryakov, K.I. Simov, Ontologically supported semantic matching, in: Proceedings of NODALIDA'99: Nordic Conference on Computational Linguistics, Trondheim, 1999.
[12] E. Vorhees, Using WordNet for text retrieval, in: WordNet: An Electronic Lexical Database, MIT Press, 1998.
[13] J.M. Whaley, An application of word sense disambiguation to information retrieval, Dartmouth College Computer Science Technical Report PCS-TR99-352, 1999.
[14] M. Lafourcade, V. Prince, Synonymies et vecteurs conceptuels, in: Proceedings of TALN'2001, 2001, pp. 233–242.
[15] M. Perkovitz, O. Etzioni, Towards adaptive web sites: Conceptual framework and case study, Artificial Intelligence 118 (2000) 245–275.
[16] M. Perkovitz, O. Etzioni, Adaptive web sites: Conceptual cluster mining, in: Proceedings of the IJCAI-99, 1999, pp. 264–269.
[17] Y. Tang, Y. Zhang, Personalized library search agents using data mining techniques, in: Proceedings of the BISC International Workshop on Fuzzy Logic and the Internet, 2001, pp. 119–124.
[18] O. Zamir, O. Etzioni, Grouper: A dynamic clustering interface to web search results, in: Proceedings of the WWW8, 1999.
[19] P. Poincot, S. Lesteven, F. Murtagh, Comparison of two document similarity search engines, in: Library and Information Services in Astronomy III ASP Conference Series, vol. 153, 1998.
[20] F. Crestani, G. Pasi, Soft Computing in Information Retrieval: Techniques and Applications, in: Series Studies in Fuzziness, Physica Verlag, 2000.
[21] H. Kraft, F.E. Petry, B.P. Buckles, T. Sadavisan, Genetic algorithms for query optimization, in: E. Sanchez, T. Shibata, L.A. Zadeh (Eds.), Information Retrieval: Relevance Feedback, Genetic Algorithms and Fuzzy Logic Systems, Word Scientific, Singapore, 1997.
[22] C. López-Pujalte, V.P. Guerrero, F. Moya, A test of genetic algorithms in relevance feedback, Information Processing and Management 38 (6) (2002) 793–805.
[23] L. Zadeh, Letter to the members of the BISC group, in: Proceedings of the BISC International Workshop on Fuzzy Logic and the Internet, 2001.
[24] D.A. Buell, D.H. Kraft, Threshold values and bolean retrieval systems, Information Processing and Management 17 (1981) 127–136.
[25] D. Choi, Integration of document index with perception index and its application to fuzzy query on the Internet, in: Proceedings of the BISC International Workshop on Fuzzy Logic and the Internet, 2001, pp. 68–72.
[26] E. Herrera-Viedma, An information retrieval system with ordinal linguistic weighted queries based on two weighting elements, International Journal of Uncertainty Fuzziness and Knowledge-Based Systems 9 (2001) 77–88.
[27] E. Herrera-Viedma, E. Peis, Evaluating the informative quality of documents in SGML format from judgements by means of fuzzy linguistic techniques based on computing with words, Information Process and Management 39 (2) (2003) 233–249.

[28] G. Bordogna, G. Pasi, Linguistic aggregation operators in fuzzy information retrieval, International Journal of Intelligent Systems 10 (2) (1995) 233–248.
[29] G. Bordogna, G. Pasi, Flexible representation and querying of heterogenous structured documents, Kibernetica 36 (6) (2000) 617–633.
[30] L. King-ip, K. Ravikumar, A similarity-based soft clustering algorithm for documents, in: Proceedings of the Seventh International Conf. on Database Sys. for Advanced Applications, 2001.
[31] M.J. Martin-Bautista, M. Vila, D. Kraft, J. Chen, User profiles and fuzzy logic in web retrieval, in: Proceedings of the BISC International Workshop on Fuzzy Logic and the Internet, 2001, pp. 19–24.
[32] D. Widyantoro, J. Yen, Incorporating fuzzy ontology of term relations in a search engine, in: Proceedings of the BISC Int. Workshop on Fuzzy Logic and the Internet, 2001, pp. 155–160.
[33] S. Miyamoto, Fuzzy sets in information retrieval and cluster analysis, Kluwer Academic Publishers, 1990.
[34] S. Miyamoto, Information retrieval based on fuzzy associations, Fuzzy Sets and Systems 38 (2) (1990) 191–205.
[35] T. Takagi, M. Tajima, Proposal of a search engine based on conceptual matching of text notes, in: Proceedings of the BISC Int. Workshop on Fuzzy Logic and the Internet, 2001, pp. 53–58.
[36] R. Ohgaya, T. Takagi, K. Fukano, K. Taniguchi, Conceptual fuzzy sets- based navigation system for Yahoo!, in: Proceedings of the 2002 NAFIPS annual meeting, 2002, pp. 274–279.
[37] S. Fernandez, A contribution to the automatic processing of the synonymy using Prolog, PhD Thesis, University of Santiago de Compostela, Spain, 2001.
[38] J.A. Olivas, Contribution to the experimental study of the prediction based on Fuzzy Deformable Categories, PhD Thesis, University of Castilla-La Mancha, Spain, 2000.