Can TF-IDF and Fuzzy Logic Improve Onomasiological Inference Ranking? Or Keywords Frequency is Good Enough?

ALBERTO BARRÓN-CEDEÑO
Universidad Nacional
Autónoma de México
Engineering Institute
Torre de Ingeniera, C. U.
MEXICO
alberto@pumas.ii.unam.mx

GERARDO SIERRA
Universidad Nacional
Autónoma de México
Engineering Institute
Torre de Ingeniera, C. U.
MEXICO
gsierram@ii.unam.mx

NICOLAS KEMPER
Universidad Nacional
Autónoma de México
Applied Sciences and
Technological Development Centre
CCADET, Ciudad Universitaria
MEXICO
kemper@servidor.unam.mx

Abstract: Onomasiological dictionaries are a simplified version of question answering systems when the user does not remember a term but its definition. Then, the query for this kind of dictionaries is a definition, in natural language, and the output is a set of the terms related to it. On this work we have taken a previously proven onomasiological inference algorithm and we have tried to improve its ranking stage by combining different weighting techniques like simple frequency counting, Fuzzy Logic, and TF-IDF.

Key–Words: Onomasiological search, ranking methods, computational lexicography

1 Introduction

An Onomasiological Dictionary (OD) is useful to identify the most appropriate word already in our lexicon to express a particular idea [4]. The purpose of an OD is to help users find out a term by means of the idea they have about it. The knowledge of a concept is highly variable and the properties that identify a particular term are quite numerous. This explains the fact that people can describe a concept in different ways by using different word sets freely organized.

The output of an OD is normally a list of terms ordered according to relevance criteria. This list must include, as it is expected, the term the user looks for. Our interest in this paper is then to improve the outputs ranking stage of an OD.

For these kind of applications, search engines based on keywords are generally implemented with Boolean techniques [6]. In other cases, a weight is assigned to each related keyword based on association functions and even word expansion is considered to get better results [7].

We have previously proposed the implementation of an OD for Spanish [8], named DEBO, on the area of Natural Disasters, and we have identified around 1000 keywords, which have been ordered on semantic paradigms through expert's knowledge. This OD is the base for this work and is described on Section 2.

Our objective is to improve DEBOs ranking process based on three main approaches: insertion of keyword occurrence (which is the base for the other three approaches), exploitation of TF-IDF, use of Fuzzy Logic, and the combination of these last approaches.

Following with the sections included on this work, section 3 describes the modifications to the ranking process of this system, section 4 shows the evaluations made and finally section 5 contains conclusions and future work.

2 The Base Inference Motor: DEBO

As we have already said, this work is based on DEBO which means *Diccionario Electrónico de Búsquedas onomasiológicas* [9] (Electronic Dictionary for Onomasiological Searching), an OD prototype developed for the area of Natural Disasters.

In order to facilitate the description of DEBO, let us propose a hypothetical user called *Hypo*. Unfortunately *Hypo* has forgotten the term that she needs to include on her report of a college excursion to the beach, but she remembers the description of the phenomenon that her teacher told her: *Great waves, produced by earthquakes in the sea, whip the coasts*. This string is the DEBOs query Q_1. The process followed by the system after *Hypo* inserts her definition is:

1. **String reception**. Typesetter signs, dots and commas are all discarded. Each letter is turned to

Table 1: Terms generated by Q_2

Leaders	Terms	Quorum
wave	seaquake	2
Sea	seaquake	
operation	accident	1
earthquake	volcanism	1

uppercase. Orthographic accents (they are used in Spanish as in *algún* [some]) are also discarded.

2. **Keywords identification**. Because we only need those keywords characterizing a specific term (keywords were previously found and stored), function words (prepositions and articles, for example) are discarded. After this step, significant meaningful words on *Hypos* concept become Q_2 [*wave, coast, earthquake, produced, sea*].
3. **Semantic expansion**. Every keyword is semantically expanded into a paradigm (a close set of keywords with common semantic characteristics that could be used in the same context). Searching follows the paradigm's leader. Once every leadership has been determined, it is searched on keywords database.
4. **Output**. There is a high possibility that the triggered paradigms produce more than one candidate term in the output. This fact explains the need to have a ranking process for the output list.

In the original DEBO implementation, ranking process is based on Boolean searching and Quorum function [1], so the weigtht of the paradigm leaders triggered by the set of keywords becomes $W_p = 1$. In *Hypos* example, Q_2 produces the output in Table 1.

The searched term is *seaquake*, which has the highest quorum in the output. On this case Quorum has worked correctly, but it does not always happen. We consider that the Quorum-based ranking process can be improved and we have made some tests using four approaches, which are described in Section 3.

3 Ranking Process Modifications

We have made four tests to improve DEBOs ranking process. The main idea of the first approach (3.1) is that if more than one keyword belongs to the same paradigm, this paradigm should be more significant than those having only one keyword. Based on this idea we have proved three more options, based on TF-IDF (3.2), Fuzzy Logic (3.3) and a combination of TF-IDF and Fuzzy logic (3.4), to determine the relevance of the keywords on each paradigm.

Table 2: Paradigms *contaminant* and *atmosphere*

contaminant			atmosphere
carbon	gas	pollution	atmosphere
combustion	harmful	soil	air
contaminant	noise	toxic	heaven
corrupt	ozone	trash	space
dioxide	smoke	waste	

Assigning a weight to each keyword based on semantic relevance could be really expensive; an expert should assign one by one. We have instead chosen to weight keywords based on their relevance inside of each paradigm. Four approaches weighting keywords are described in the next sections and comparisons of the obtained results are shown in Section 4.

3.1 Weighting paradigms based on keywords frequency Paradigm

The first approach is simply based on frequency. See this query: *it is to soil the environment, like air with toxic gases or excessive noise. The water with grease, detergents, trash and toxics. The ground with pesticides, grease, etc.*. The correct output should be contamination. We will concentrate in two paradigms included in Table 2, where the leader paradigm and the keywords there contained are shown.

Definition keywords *toxic, trash, gas, noise* and *soil* are included in the paradigm *contaminant*, while *air* belongs to paradigm *atmosphere*. Using Quorum, paradigms *contaminant* and *atmosphere* have exactly the same relevance in the ranking process no matter that one of them have more related keywords.

Since the purpose is to consider the number of query keywords included in a paradigm to define its relevance, we do not sum 1 to the paradigm-related term, but now we compute $W_p = \sum_{k_p} c(k)$ where $c(k)$ represents how many keywords of the query belong to the paradigm.

Table 3 shows how this calculation method affects the output[1]. Here, *contamination* appears higher in the list ordered by Frequency. This method is the base for the next three ones, described below.

3.2 Weighting Keywords with TF-IDF

As we have seen in 3.1, weights can be assigned according to the keyword frequency in all paradigms. After that, a brief modification on that approach has

[1] f. d. s. means flight of dangerous substances

Table 3: Quorum and freq. methods comparison

Quorum		Frequency	
plague	3	poisoning	8
hail storm	2	**contamination**	7
drought	2	drought	6
collapse	1	plague	3
contamination	1	hail storm	2
earthquake	1	collapse	1
f. d. s.	1	earthquake	1
flood	1	f. d. s.	1
frost	1	flood	1
hurricane	1	frost	1
poisoning	1	hurricane	1
...			

Table 4: Quorum and TF-IDF methods comparison

Quorum		TF-IDF	
drought	2	drought	66.34
hail storm	2	**flood**	**35.40**
collapse	**1**	rain	30.34
earthquake	1	seaquake	27.71
flood	1	hail storm	21.30
frost	1	earthquake	20.22
plague	1	volcanism	18.54
rain	1	collapse	17.55
seaquake	1	plague	15.48
s. of slope	1	s. of slope	1.98
volcanism	1	frost	1.68

been made: the use of TF-IDF [5] to differentiate the relevance of a keyword in a set of paradigms.

It is possible to define the relevance of a keyword in every paradigm inside an entire universe of them, so the fewer keywords appear on the universe of paradigms, the greater their weight. The original TF-IDF is reproduced in Formula 1.

$$w_{ik} = \frac{tf_{ik} * log(\frac{N}{n_k})}{\sqrt{\Sigma_{j=1}^{t}(tf_{ij}^2)(log(\frac{N}{n_j}))}} \quad (1)$$

where tf_{ik} is the frequency of occurrence of term t_k in document D_i, N is the size of the documents collection, and n_k is the number of documents in the collection having the term t_k.

The longer a document is, the higher the retrieval probability is. The summation in the denominator is included to normalize the values. However, this denominator is designed to deal with long documents, but in this case paradigms are composed only by a small set of keywords, so it is possible to reduce formula 1 to formula 2.

$$w_{ik} = tf_{ik} * log\left(\frac{N}{n_k}\right) \quad (2)$$

So, for example, *tectonic* is an important keyword because it appears in only two paradigms (crust and plate) meanwhile *water*, belongs to nine different paradigms: *water, environment, phreatic, drops, ice, humidity, rain, sea* and *nature*. This means that many terms will probably be related to these paradigms; so *water* should not be too important. For that reason, the weights assigned to these keywords by TF-IDF are $w_{tectonic} = 1.98677$ and $w_{water} = 1.03253$ respectively.

The relevance of every paradigm is the summation of the weights of the keywords that it contains. Then, a summation of the paradigm weights (as in Section 3.1) is made to determine the probability of how each term retrieved is the searched one.

See the query *natural effect pertaining to a stage of the cycle of the water, when the water of the terrestrial crust evaporates, and is concentrated in the stratosphere for, after having a given temperature, hurries in drops of water*. The output generated through Quorum and TF-IDF approaches is shown in Table 4.

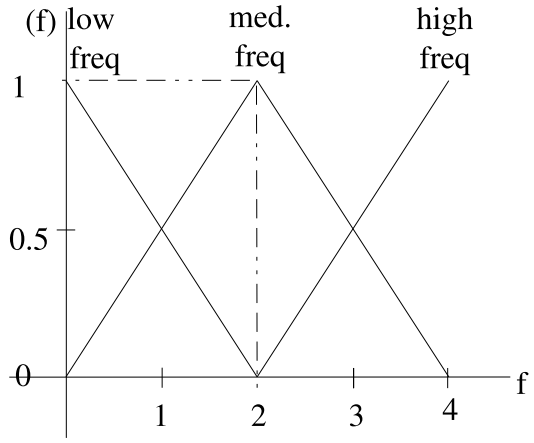
3.3 Adding Fuzzy Sets to the Ranking Process

Fuzzy Logic [11] has been already used in NLP and Information Management tasks, as well as in Hierarchical Clustering [2], Management Systems [10], and in the implementation of dictionaries of synonyms and antonyms [3]. In our case, it is thought that through Fuzzy Sets the rigidity of Boolean searching could be avoided and the weight of each triggered paradigm can be determined with more precision.

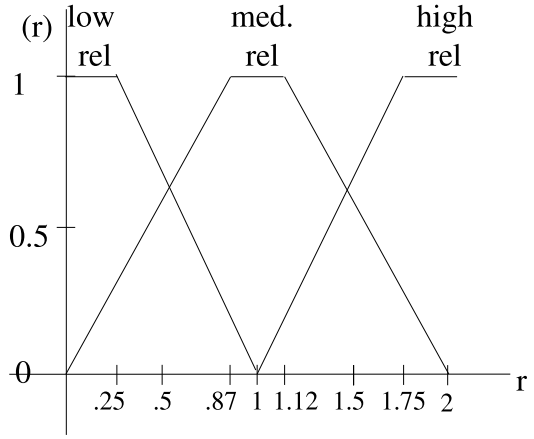
We have implemented Fuzzy Sets to determine the strength of a paradigm based on keyword occurrence in the query just as in section 3.1. Thus, Fuzzy Sets have been empirically designed for the frequency of keyword occurrence in a paradigm. The membership function is called Frequency; functions and graphical representation are shown in figure 1 (a).

Fuzzy output is given by a Singleton function, (figure 1 (c)), the insertion of very strong relation is justified in 3.4. The Fuzzy Rules are shown below:

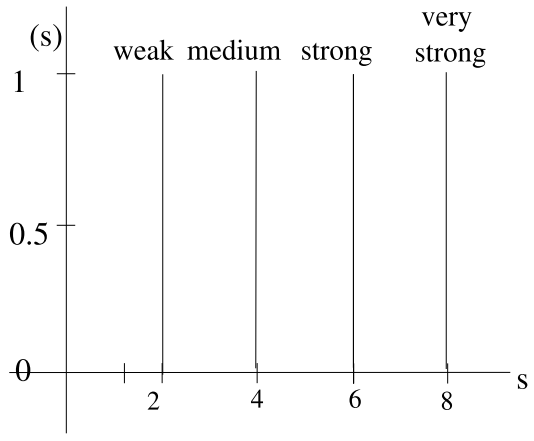
$l_{term} = weak$ `if` $freq = low$
$l_{term} = med$ `if` $freq = med$
$l_{term} = strong$ `if` $freq = high$



(a) keywords to paradigms (frequency)



(b) keywords and its paradigm (occurrence)



(c) Fuzzy output

Figure 1: Fuzzy sets

Table 5: Quorum and Fuzzy Methods Comparison

Quorum		Fuzzy	
rain	2	rain	7.952
hail storm	2	drought	7.952
drought	2	collapse	3.976
collapse	1	f. d. s.	3.976
f. d. s.	1	flood	3.976
flood	1	plague	3.976
plague	1	seaquake	3.976
seaquake	1	sliding of slope	3.976
sliding of slope	1	**snowfall**	3.976
snowfall	1	hail storm	3.952

In the case of the query *atmospheric phenomenon in which the water in clouds falls exclusively in ice form*, the user was looking for the term *snowfall*. The outputs sorted by the Quorum and Fuzzy approaches are compared in Table 5[2].

3.4 Combinig Fuzzy Sets and TF-IDF

Finally, we have decided to combine the methods described in 3.2 and 3.3. As we have previously said, the inclusion of more than one keyword from the query in one paradigm becomes more relevant for term search. On the other hand, the rigidness of natural numbers could be considered unsafe for the sorting process, good reason for applying Fuzzy Logic. A membership function for the TF-IDF values, called Relation, has been added and determines the relevance of a paradigm based in the keywords there contained. The fuzzy sets for this function are included in figure 1 (b). By including the new membership function occurrence, nine new rules have been designed:

$l_t = wea$ `if` $rel = low \wedge freq = low$
$l_t = med$ `if` $rel = med \wedge freq = low$
$l_t = strong$ `if` $rel = high \wedge freq = low$
$l_t = weak$ `if` $rel = low \wedge freq = med$
$l_t = med$ `if` $rel = med \wedge freq = med$
$l_t = strong$ `if` $rel = high \wedge freq = med$
$l_t = med$ `if` $rel = low \wedge freq = high$
$l_t = vStrong$ `if` $rel = med \wedge freq = high$
$l_t = vStrong$ `if` $rel = high \wedge freq = high$

See the process in the next example (shown in Spanish). Now, *Hypo* has a new query: $Q_1 = movimientodelatierra$ (earth movement). After the keywords identification, $Q_2 = [tierra, movimiento]$. During the semantic expansion, $s \in Q_2$ generate the lemma array $l =$

[2] f. d. s means “flight of dangerous substances”.

Table 6: Paradigm leaders for the query lemmas

lemma	TF-IDF	Paradigm leader
mov	1.38	derrumbe (collapse)
		reacomodo (rearrangement)
		desacomodo (disarrangement)
		rotación (rotation)
movimiento	1.68	migración (migration)
		naturaleza (nature)
tierra	1.38	ambiente (environment)
		naturaleza (nature)
		placa (plate)
		suelo (ground)

$[movimiento, mov, tierra]$ which is related to the paradigm leaders shown in Table 6.

Since there are 97 paradigms in total, the TF-IDF for mov string is 1.3847 as it is shown below:

$$w_{[derrumbe][mov]} = 1 * log\left(\frac{97}{4}\right) = 1.3847.$$

This value is true for the four paradigms containing the string *mov*. The fuzzyfication stage starts here. Only one case for the Frequency Set case is shown due to space. Lemmas *movimiento* (movement) and *tierra* (land) belong to the *naturaleza* (nature) paradigm, so $f_{naturaleza} = 2$ and belongs to the medium frequency set with a membership value of 1 (see figure 1 (a)).

After making the fuzzyfication process for every keyword (Occurrence) and paradigm (Frequency) the corresponding rules are triggered, and after de defuzzyfication process, the output is obtained (Table 7 with the candidates in English). The rules corresponding entirely to *Hypos* query are in bold.

These four experiments have been made for the ranking process of the output in DEBOs inference engine.

4 Evaluation

Since the inference algorithm has not been modified and we have worked only on the ranking process, we do not include an evaluation in terms of Precision and Recall of the different approaches. If so, it will be the same for all of them. Instead, the approaches have been evaluated on the basis of the searched term position in the output list of each one, considered the aspect to be improved.

Table 7: Output for Q_1

Rank	Candidate	Weight
1	**earthquake**	2.6343
2	explosive growth of the population	1.8318
3	**tidal wave**	1.8318
4	**collapse**	1.3171
5	drought	1.3171
6	hurricane	1.3171
7	plague	1.3171
8	**sliding of slope**	1.3171

Table 8: Position of the searched term in the output

P	Q	Q+F	T	F	T+F
1	78.57	76.78	55.35	67.85	75.00
2	12.5	16.07	21.42	8.92	8.92
3	3.57	1.78	3.57	3.57	3.57
4	5.35	5.35	10.71	3.57	5.35
5			5.35	1.78	
other			3.57	14.28	7.14

It is necessary to say that, for the sake of tests and evaluations, some students of undergraduate programs of Engineering and Linguistics have been asked to give the description of some of the terms considered in this work. Out of a total of seventy-four definitions given, fourteen were discarded because of ambiguity (as in the example of water falls from clouds and some thunders are produced that points to rain and/or storm), resulting in a set of sixty-five test queries.

Of the sixty-five test queries tested, fifty-six included the searched term in the output list. The comparison of the five approaches through these queries is shown in figure 3 when the y axis represents the position of the searched term.

Considering that the best position is 1, Quorum method shows the searched term in this position in 44 of the 56 experiments, while Frequency shows it the first place in 43. Percentage for each position using quorum, quorum and frequency, TF-IDF, fuzzy, and TF-IDF and fuzzy are shown in Table 8.

It seems that the simple Quorum method is definitely better than the other approaches. Nevertheless, see this query: *effect of wear produced by water or air depending on the place*.

The searched term is *erosion* and the output using Quorum and TF-IDF methods is shown in table 9.

Table 9: Quorum and TF-IDF granularity comparison

Quorum		TF-IDF	
term	weight	term	weight
drought	2	drought	3.80
accident	1	accident	1.98
collapse	1	snowfall	1.98
desertification	1	desertification	1.68
erosion	1	erosion	1.68
flood	1	hurricane	1.38
hail storm	1	radiation	1.38
hurricane	1	collapse	1.03
plague	1	flood	1.03
radiation	1	hail storm	1.03
...			

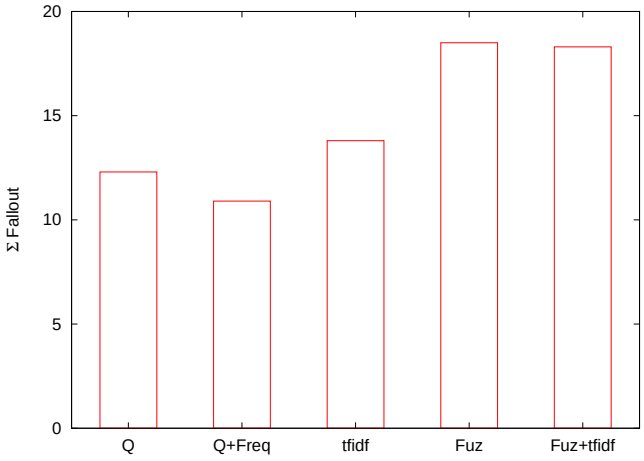


Figure 2: Fallout Comparison

erosion appears in the second position of the Quorum-based output list, but it also appears with other twelve terms in the same position. Because Quorum generates an output divided in subsets, where the probability of containing the searched term decreases from set to set in a discrete way, the position of the searched term could be false. In the TF-IDF-based output list, erosion appears in the fifth place, but only one more term has the same weight. This is only an example of a phenomenon that has occurred in an important section of the tests.

On the other hand, not only in the case of TF-IDF but in the other approaches, the ranking method does not divide the output list in subsets. The four methods are based on the individual competition of a term to occupy the top of the list, and this is measured through a continuous scale.

Then, a further evaluation has been made. Fallout determines the precise position of the searched term into the output list, no disregarding the numerical value assigned to it. Thus, Frequency method gives better results as compared to the other approaches, which is shown in figure 2.

Furthermore, we can see that although TF-IDF has not obtained bad results, the fuzzy approach has obtained the worst ones, even after improve it with the combination of fuzzy and TF-IDF approaches.

5 Conclusions and Future Work

In this paper four approaches have been shown, in addition to Quorum Method, for ranking the output given by an Onomasiological Dictionary in the area of Natural Disasters, named DEBO.

Our experiments have considered various resources for paradigm weighting based on the keywords of the query contained in them. In all four cases, each keyword in the query adds up points to the total weight of the paradigms looking for term candidates. For this summation, it has been considered to score keywords with four approaches:

1. The occurrence of a keyword gives one point to every paradigm it belongs to, so the weight of each paradigm is the summation of its keywords included in the query (Frequency).
2. The relevance of each keyword in a paradigm is defined upon a simplified version of TF-IDF (Formula 2) with respect to the universe of paradigms. This score is added to each paradigm containing a given keyword. This is the reason why this method has been named TF-IDF.
3. The weights of method a) have been taken in a fuzzy manner. We have just tried to smooth the position differences in the output
4. A combination of 2 and 3 approaches.

The results obtained with the use of the four approaches above described have given interesting conclusions. One of them refers to the fact that a simple change to the Quorum method (to weight each triggered paradigm with the number of query-keywords on it instead of assigning 1) has improved the output rank.

The second conclusion is related to the question stated in the title of this paper. It has been observed that the application of TF-IDF and Fuzzy Logic in the ranking process becomes more complex and the ranking results are worse than the Quorum process. For this reason, the frequency approach (Section 3.1) is simple and good enough for the application here described.

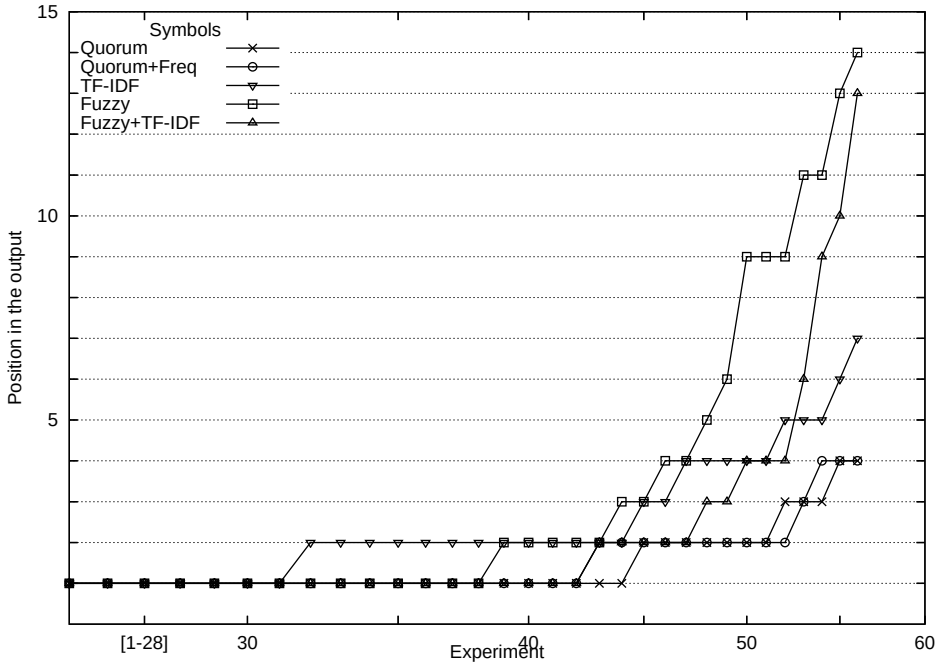


Figure 3: Comparison of the five approaches

Future efforts will be addressed to the creation of a method to populate automatically and weigh the paradigms and keywords of our database, so that new terms and definitions without human intervention may be added.

Aknowledgementes

This paper has been supported by the National Council for Science and Technology (CONACYT) of Mexico, Ref. 46832-H; the General Direction of Graduate Studies (DGEP), UNAM; and the Macro Project named Tecnologas para la Universidad de la Informacin y la Computacin, UNAM.

References:

[1] C. Cleverdon, *Optimizing convenient on-line access to bibliographic databases*, in *Information Services and Use*, 4 (1), 1974, pp. 37–47.

[2] Yih-Jen Horng, Shyi-Ming Chen, Yu-Chuan Chang and Chia-Hoang Lee, A New Method for Fuzzy Information Retrieval Based on Fuzzy Hierarchical Clustering and Fuzzy Inference Techniques, in *IEEE Transactions on Fuzzy Systems*, Vol. 13, No. 2, 2005.

[3] S. Lanza, J. Graa and A. Sobrino, Introducing FDSA (Fuzzy Dictionary of Synonyms and Antonyms): Applications on Information Retrieval and Stand-Alone Use, in *Mathware Soft Comput. 10*, No. 2-3, 2003, pp. 57-70.

[4] F. W. Riggs, Ethnicity: Intercocta glossary, *Honolulu, University of Hawaii*, 1985.

[5] G. Salton and M. J. McGill, Introduction to modern information retrieval, McGraw-Hill, New York, 1986.

[6] G. Salton, *The Use of Extended Boolean Logic in Information Retrieval*, SIGMOD Record, Vol.14, No.2, pp. 277-285, 1984.

[7] G. Sierra and J. McNaught, *Design of an onomasiological search system: A concept-oriented tool for terminology*, in Terminology. Num.1, Volume 6, 2000.

[8] G. Sierra, Bases para la bsqueda onomasiolgica de trminos, *Universidad Nacional Autnoma de Mxico*, 1996

[9] G. Sierra, *The Onomasiological Dictionary: A Gap in Lexicography*, in *Proceedings of the Ninth EURALEX International Congress*, Stuttgart, 2000, pp. 223–235.

[10] P. Subtil, N. Mouaddib and O. Foucoat, A Fuzzy Information Retrieval and Management System and its Applications, on *Proceedings of the 1996 ACM symposium on Applied Computing*, Philadelphia, 1996, pp. 537–541.

[11] L. A. Zadeh, Fuzzy sets, Inf Control 8, 1965, pp. 338–353.