A fuzzy logic system for classifying the contents of a database and searching consultations in natural language

Ariel Gómez[1], Jorge Ropero[1] and Carlos León[1], Member, IEEE.
[1] Department of Electronic Technology, University of Seville, Spain.
ariel@us.es , jropero@dte.us.es, cleon@us.es

Abstract—**This paper presents a method for the classification of the contents in a database in order to answer to user consultations using natural language. Artificial Intelligence (AI) is used to relate these consultations to the database contents. The system is based on a fuzzy logic engine to take advantage of its so suitable properties for this application and is ideal for sets of accumulated knowledge that can be built in hierarchic levels by a tree structure. The eventual aim of this system is the implementation of a virtual web assistant for an internet portal.**

Index Terms— Artificial Intelligence, Fuzzy Logic, database, searching.

I. MOTIVATIONS

The access to the contents of an extensive set of accumulated knowledge – a database, a summary of documents, web contents, etc – is becoming an important concern in the last decades. In some occasions, a user tries to get to them knowing that what he needs is out there but ignoring the exact denomination for what he is looking for and/or the suitable method to make the extraction of the desired knowledge. All these disadvantages meet increased when the user in question is not a habitual of the matter or there are ambiguous contents, bad organization or, simply, complex topics or a great amount of information difficult to manage.

In these cases the unsuccessful attempts can turn out to be frustrating for not using the exact term to make the consultation - a machine only will answer adequately if it is asked in an exact way -, and one can eventually end in a paradox: the less one knows the more difficult it is to find the answers.

In many cases the solution is to look for another person who is an expert on the topic. Actually, the demanded help is an interpreter who is able to generate a syntactically and semantically correct search that leads us to the obtaining of the desired answers. Consequently, the need of an assistant who interprets the vague information we have arises, giving us concrete answers related to existing contents in its database. This should be based on an estimation of the certainty of the relation between what we have expressed in natural language and the contents stored its database.

To solve this, we develop a method of classification of the contents by means of the creation of a few indexes based on key words, and a method of consultation based on a fuzzy logic application provided with an interface that one may interact with in natural language. We propose then an application of the artificial intelligence (AI) based on the use of fuzzy logic.

II. FUZZY LOGIC

Fuzzy logic arises as response to the inflexibility of the classic binary logic. In this one, the middle term does not exist: either it is hot or it is cold. Fuzzy logic gives us the possibility of having intermediate grades: cool, fresh, warm… and, besides, by means of a set of functions, a degree of flexibility may be given to these epithets: what may be cool for a Sevillian might be mild for a Berliner.

Fuzzy logic turns out to be useful for:

a) Dealing with uncertainty.
b) Dealing with precise information held together with uncertainty.

Therefore, we have simultaneously precise information and uncertainty. Using fuzzy logic certain quantity of precision is sacrificed in favour of uncertainty with the hope of obtaining conclusions that, though more vague, are more robust. [1-6]

III. MODE OF OPERATION OF THE METHOD OF CLASSIFICATION AND INFERENCE

A. *Classification of accumulated knowledge*

We propose to group the knowledge in sets of different levels: the final content, it is to say, every element, belongs to an N level set. Several level N sets with some common features gather in groups forming another N-1 level set. Analogously, several N-1 level sets with some common features gather in other groups forming an N-2 level set, and so on up to coming to level 0 set (N-N), which represents the totality of the accumulated knowledge. The proposed classification is seen in figure one, clearly forming a tree structure.

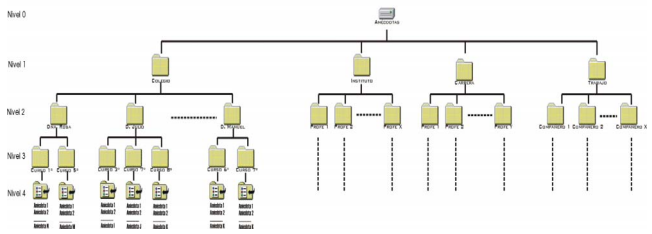


Figure 1 - Classification in a tree structure.

The first step of the processing consists of distinguishing the words of the consultation. These words must be searched in a database which must contain the words that are related somehow to the content of the matter we have to deal with. Another database with the possible answers is necessary.

Key words are assigned to every possible answer in order to identify it. These words are chosen among those that could appear in a possible consultation. The belonging of these words to every level is determined by means of a few numerical coefficients indicating how significant the considered word inside the level in question is. It is important to notice that the same word can belong to several different sets.

B. *Example*

To illustrate the process, the following example is used: a user of the University of Seville internet portal asks 'How can I find a professor e-mail address?' From here we would extract: find, professor, e-mail, and address. We could also add some related words: get – related -, professors – plural -, electronic – related - and mail - similar.

The following step is the assignment of the coefficients to every word. Sure that there are many questions relative to find something so this word does not have a very high coefficient; there must be enough questions relating to professors, too; with regard to e-mail or address it is sure that there are fewer questions so that these words must have higher coefficients than others.

This way, coefficients for N level - level 3 in the example – might be:

E-mail = mail = 0.7.
Address = 0.6
Professor = professors = 0.5.
Find = get = 0.3.

Due to the use of fuzzy logic it is not necessary to be very precise with indexes, the system tends to be convergent and to identify the content. When identification tests for every anecdote are made all key words are introduced and it is decided whether it is necessary to change these coefficients or not.

Then we must group all the questions according to the section they belong to form level N-1 groups. The word indexes will not be the same but they will reflect how significant is this word in the following level. Likewise, the top level tables would be built up to coming to a level 0 table in which all the words would appear. This way the whole content of the database would be grouped in hierarchic levels, identified by a set of key words with an assigned index of importance which means how much this word identifies the content of the knowlege database as belonging to one of the levels.

As mentioned above, the aim of the system is to allow the user to formulate consultations by means of questions in natural language, to relate them to the contents using the index generated with fuzzy logic and to obtain the answers corresponding to the contents stored in the set of knowledge.

C. *Fuzzy logic system*

The inference method for finding out the knowledge demanded by the consultation is described in this section.

All the key words are extracted for comparison with the ones contained in our word database.

The belonging to every level 1 set is analyzed bearing in mind the value returned by the fuzzy engine. If the level of certainty is lower than a predefined value, the belonging of the content to the corresponding set is rejected. The facts of starting from level one and using a tree structure make possible the rejection of a great amount of content, which will not be considered in future searches.

For every set that has overcome the minimum certainty

threshold, the process is repeated and the coefficients of belonging corresponding to every level 2 set are evaluated. Again, the sets in which the returned by the fuzzy engine certainty does not overcome a certain minimal threshold are rejected. If they overcome the threshold, the method for determining the belonging to the following level is applied to them. This process is repeated up to coming to the last level. The answers correspond to those last level elements which certainty overcomes the definite threshold. There can be more than one answer. The vaguer are the questions, the more answers we will obtain.

The heart of the fuzzy logic system is the fuzzy engine. This engine is the responsible of determining the probability for the key words contained in the consultation of belonging to a certain fuzzy set in a concrete level. The engine must evaluate the belonging to every set for the corresponding level. For that reason, the engine takes the coefficients of the key words for that set as inputs. The fuzzy engine output will be determined by the defined rules. These rules are of the IF … THEN type. An example of rule might be the following one:

IF word_index1 is HIGH AND word_index2 is MEDIUM AND word_index3 is LOW, THEN output is HIGH.

The definition of these rules corresponds to the system administrator.

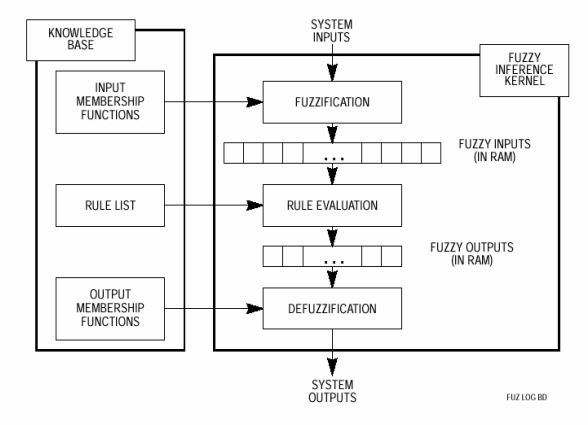


Figure 2 - Fuzzy logic system

D. System administrator

In previous sections the need of a system administrator has been mentioned. The functions of this administrator must be basically three.

a) Defining and modifying the coefficients of belonging to a topic, paragraph or answer for a certain word.

b) Adding new words to the database when necessary.

c) Making a system feedback asking the eventual users their opinion about the answers given by the assistant in order to take the necessary steps in every case.

IV. TESTS AND RESULTS

A. System simulation

To make the first functioning tests for the proposed method, we used a knowledge database consisting of a set of the more frequent questions belonging to the administrator system. Concretely, the analyzed case considers 133 questions. This allows to study the functioning of the method on a population big enough to get results which are reliable and comfortable to handle at the same time.

Knowledge is structured in three levels: Level 1 is correspondent to the Topic to which the question belongs; level 2 corresponds to a Section inside every topic – the number of Sections for each Topic varies between 3 and 10 - ; finally, level 3 corresponds to every concrete Question inside every Section in every Topic - the number of Questions for each Section also varies between 1 and 10).

For every Question, between 3 and 6 key words are defined. Nevertheless, the user will include in his consultation from 1 to 6 of these words. Defining an only engine with a few inputs causes the rapid saturation of the system, whereas to define an only six input engine causes values with a very low degree of certainty. The solution provided is to implement a flexible system with variable inputs. If the user consultation three key words as much, a three input fuzzy engine is used, whereas if the user consultation includes four or more key words, the system will use a five input fuzzy engine.

As said above, rule definition corresponds to the administrator. Logically, the more inputs the engine has, the more rules there are. For example, for a three inputs engine, the inputs can take three values: LOW, MEDIUM and HIGH. The outputs can take the values LOW, MEDIUM-LOW, MEDIUM-HIGH and HIGH. The inference rules defined are:

If all inputs are LOW, output is LOW.

If one input is MEDIUM and the others are LOW, output is MEDIUM-LOW.

If two inputs are MEDIUM and the other are LOW, output is MEDIUM-HIGH.

If all the input are MEDIUM or one input is HIGH, output is HIGH.

The possible combinations generate 27 rules for the fuzzy engine. A five input engine generates 243 rules.

All the obtained results are shown in the following figures.

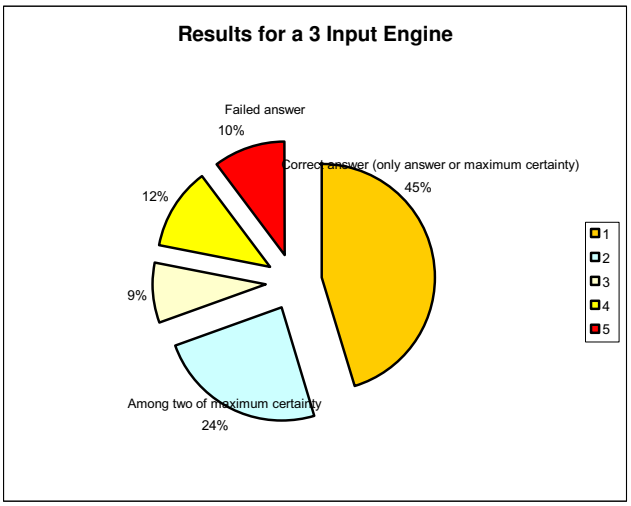


Figure 3 - Results for a 3 input engine

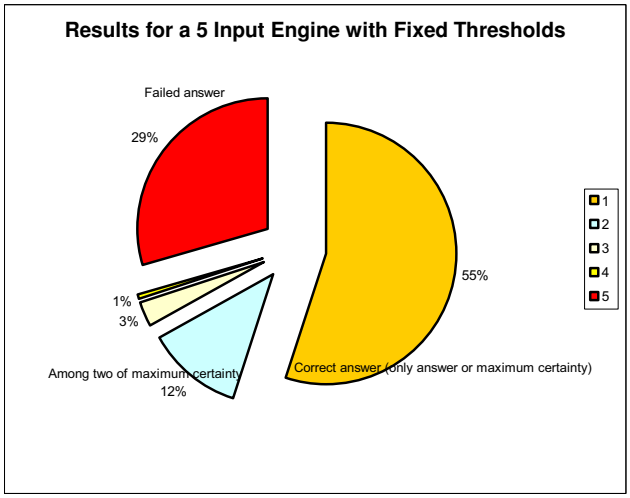


Figure 4 - Results for a 5 input engine with fixed thresholds.

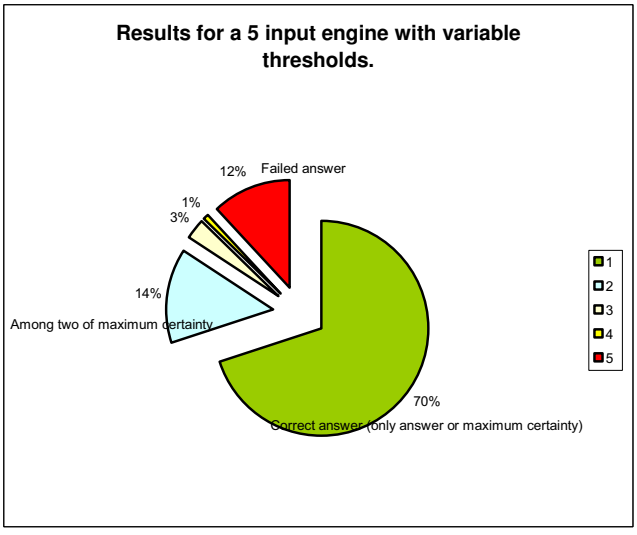


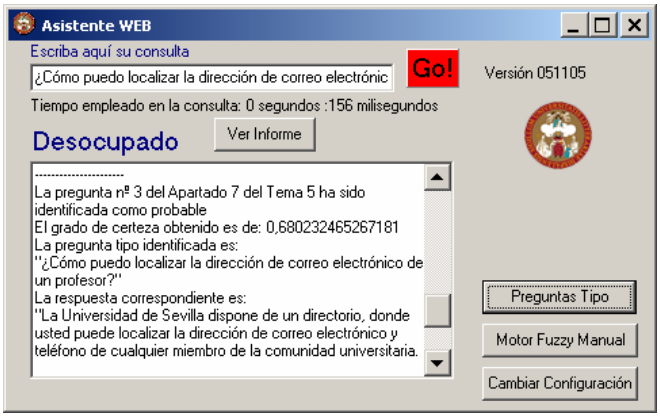
Figure 5 - Results for a 5 input engine with variable thresholds.

V. Conclusions and Future Lines of Work

Up to now, the obtained results using fuzzy logic are good enough, as the number of correctly detected consultations is high.

Future lines of work must begin by a possible improvement of the system based on tests using different parameters for the fuzzy logic system and the execution of new tests where the consultations have a different structure than those found in the database. Likewise, other words must be added to the database which contains the related words. This way, we will be able to consider synonymous or expressions that are similar to the standard one.

Another line of investigation is the creation of the user interface. Up to now we have only developed a version for the administrator, which can be seen in the following figure.



.

Figure 6 – Administrator version interface

References

[1] B. Martín del Brío, A. Sanz Molina, *Redes neuronales y sistemas borrosos*. Ra-Ma, 2001.

[2] *Fuzzy Logic Toolbox. User`s guide.* The Mathworks Inc., 2002.

[3] T. Bouaziz, A. Wolski, *Applying Fuzzy Events to Approximate Reasoning in Active Databases.* Proc. Sixth IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'97). July 1-5, Barcelona, Spain.

[4] S. A. Moriello. *Intelectos ultrarracionales*. ISSN: 1597-0223, 2004

[5] D. Xie. *Fuzzy Association Rules Discovered on Effective Reduced Database Algorithm*. FUZZ-IEEE 2005, (The IEEE International Conference on Fuzzy Systems), Reno, USA, May, 2005.

[6] P. Bedi, H. Kaur, A. Malhotra, *Fuzzy dimension to databases*. 37th National Convention of Computer Society of India, Bangalore, India, November 2002.

[7] D.Olsen. *Fuzzy Logic Control in Autonomous Robotic.* University of Minnesota. Research Project. October, 2002