Model of an intellectual search engine represented by the fuzzy sets theory

Hristina Moneva

Abstract: *In this report an attempt to explain the role and place of the fuzzy set theory with its typical techniques and methods in the process of the knowledge representation is proffered. A usage of the fuzzy sets and fuzzy logic for organizing a knowledge base is described. Their role with their typical characteristics in the process of designing a system for searching and classifying documents based on knowledge is examined. A generalized model of an intellectual system for searching and classifying documents in the Web is proposed.*

Key words: *fuzzy sets, intellectual systems, systems based on knowledge, search engines, classification.*

INTRODUCTION

When information is searched in Internet or LAN a request is send and a response is received. This is a list of documents and sometimes for the user it is hard to locate the desired knowledge. The received documents are not thematically grouped, not ordered and also it is not possible to achieve the level of adequacy that is needed.

To solve these problems a new approach for designing a search engine is proffered – an intellectual search engine based on knowledge. In its knowledge base will have data about documents, keywords that describe them and the domains that they belong to (basically). Systems, which have a kernel that is knowledge base or model of domain, are described with some high level language close to natural language. They are called intellectual [2]. With this approach, a person who publishes a few documents has to give them certain characteristics before adding them to the knowledge base. In this case they are the experts whose knowledge is stored in the intellectual system. In contrast to standard systems is that knowledge base is not static, but is getting larger all the time.

Attempting to formalize the common knowledge, the researchers faced the problem to represent adequately the vagueness in a conventional mathematical way.

The attempt to formalize the common knowledge met the problem to represent it by using conventional mathematical apparatus to solve it. In case we are dealing with a non-numerical quantity that cannot be measured against a numerical scale, than we cannot use a numerical universe. Than it is said that the elements were taken from a psychological continuum. An example of such a universe could be {large, medium, small, very small}. These characteristics usually are fuzzy and cannot be synonymously interpreted.

For that reason tasks that are solved in the intellectual systems often have to work with an uncertain knowledge which cannot be interpreted as a pure true or false (logical true/false or 0/1). There is a knowledge which reliability can be expressed with an intermediate number, for example 0.7.

How to represent such knowledge formally keeping the fuzziness and uncertainness? In order to solve these problems, the American mathematician Lotfi Zadeh proffered a formal apparatus of fuzzy logic and fuzzy arithmetic in the beginning of 70's. Later this trend marked the beginning for one of the areas of AI –soft computing [2].

Zadeh introduced one of the basic concepts in fuzzy logic – the concept of a linguistic variable. *LINGUISTIC VARIABLE* – this is a variable which values are words or sentences. The set of values it can take is called its term set. Each value in the term set is a fuzzy variable defined over a base variable. The base variable defines the universe of discourse for all the fuzzy variables in the term set. The fuzzy set is defined with some base scale B and membership function $\mu(x)$, $x \in B$ which values are into [0,1]. Consequently the fuzzy set B is a collection of ordered pairs $(x, \mu(x))$, where $x \in B$.

When the membership function gives the subjective expert level, the concrete value of base scale is the definition of a fuzzy set. This function is not a probabilistic that has an objective character and is not following any other mathematical dependencies.

The fuzzy sets give possibilities to define the subjective opinion of different individuals.

FORMULATION OF THE TASK

Let **U** be the universe of discourse (for short just universe) containing all sets $\mathbf{A_1,A_2...A_n}$ as elements. Elements of these sets are domains and sub domains that are organized in hierarchies. The set **U** is called universe, because the system pretends to be able to consist of all domains and sub domains that can come into consideration.

Let the elements of set **D** be the documents $\mathbf{d_1,d_2...d_m}$.

Let the set **R** be the binary relation from a set **U** to a set **D**, which gives information about classification of documents into domains and sub domains. The relation **R** is described by characteristic function of set **R** - $\left\|\mu_{Rij}(A,d)\right\|$, i=1,2...n, j=1,2...m, define as

$$\mu_{Rij}(A,d)=\begin{cases}1, ako(A_i,d_j)\in R\\ 0, ako(A_i,d_j)\notin R\end{cases}$$ [1].

Let **K** be the set with elements $\mathbf{k_1,k_2...k_l}$ – keywords describing some document of the set **D** with the fuzzy relation **S** defined with $\mu_S(d,k)\in[0,1]$.

Now we have described all elements of the knowledge base, that is part of an intellectual classifying and searching system.

To define how these keywords are describing all documents in a domain, specify this domain too. To define this, we have to find max-min and min-max composition of **R** and **S** (or optimistic and pessimistic expectations [4]). Now we have the fuzzy relations $\mathbf{T_1}$ and $\mathbf{T_2}$, where for T_1: $\mu_{R\circ S}(A,k)=\sup_{d\in D}\left[\min(\mu_R(A,d),\mu_S(d,k))\right]$

and for T_2: $\mu_{R\circ S}(A,k)=\inf_{d\in D}\left[\max(\mu_R(A,d),\mu_S(d,k))\right]$ [3].

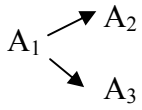
The relation $\mathbf{T_2}$ is always set of zero elements because normally always will exist a keyword that does not appear in all domains and sub domains, otherwise this keyword is not important for us.

Now all documents are classified in different domains / sub domains with the set of keywords using the fuzzy sets and the fuzzy relations theory. The domains and sub domains are characterized with keywords too.

All this can be used for searching on keywords and letting us represent the result classified in different domains / sub domains consisting of a list of documents for each of them.

For example:

The sets $\mathbf{A_2}$ and $\mathbf{A_3}$ are subsets of $\mathbf{A_1}$: $A_2\subset A_1$, $A_3\subset A_1$ or presented as oriented graph:



R	**d_1**	**d_2**	**d_3**	**d_4**	**d_5**	**d_6**
A_1	1	1	0	0	0	0
A_2	0	0	1	1	0	0
A_3	0	0	0	0	1	1

S	**k_1**	**k_2**	**k_3**	**k_4**	**k_5**	**k_6**
d_1	0,9	0,6	0	0	0	0
d_2	0,7	0	0,7	0	0	0
d_3	0	0,4	0	0,6	0	0
d_4	0	0	0	0,8	0,5	0
d_5	0,3	0	0	0	0	0,4
d_6	0	0,8	0	0	0	0,9

T_1	**k_1**	**k_2**	**k_3**	**k_4**	**k_5**	**k_6**
A_1	0,9	0,6	0,7	0	0	0
A_2	0	0,4	0	0,8	0,5	0
A_3	0,3	0,8	0	0	0	0,9

T_2	**k_1**	**k_2**	**k_3**	**k_4**	**k_5**	**k_6**
A_1	0	0	0	0	0	0
A_2	0	0	0	0	0	0
A_3	0	0	0	0	0	0

LEVEL OF ADEQUACY (LEVEL OF FUZZY RELATION)

A new concept of LEVEL OF ADEQUACY is introducing. This is a number into [0,1] which is set by the user and characterizes the relevancy between searched keyword (phrase) and proposed documents by the system.

Let the searched word be **ks**. Let the searched level of adequacy to this word be **α_{ks}**. Consequently the given level of adequacy is the level **α_{ks}** of the fuzzy relation **S** of **DxK**, respectively **$\mu_S(d,k) \geq \alpha_{ks}$**.

The results under the conditions mentioned above can be presented as a hierarchical structure which elements have this type *Domain_name[pessimistic÷optimistic_expectation]*.

Because the min-max composition is always zero and to be clear and easy visible the given level of adequacy, the result may be presented this type: *Domain_name[level_of_adequacy÷optimistic_expectation]* or $A_i[\alpha_{ks} \div \max\min \mu_{R \circ S}(A_i, ks)]$.

For example:

If the searched word is **$ks=k_1$** and the level of adequacy is **$\alpha_{ks}=0,7$**, then the result from the example mentioned above will be only the documents **d_1** and **d_2** under these conditions and can represented as:

$A_1[0 \div 0,9] \rightarrow d_1(0,9)$, $d_2(0,7)$

, where:

A_1 - this is the domain where the documents belong to and **[0÷0,9]** is the interval of which the results have to be.

If **$\alpha_{ks}=0,8$** then we have in return only one document - **d_1**. The result is:

$A_1[0 \div 0,9] \rightarrow d_1(0,9)$

.

If the result is presented as $A_i[\alpha_{ks} \div \max\min \mu_{R \circ S}(a_i, ks)]$ then we have:

$A_1[0,7 \div 0,9] \rightarrow d_1(0,9)$, $d_2(0,7)$ in the first case and

$A_1[0,8 \div 0,9] \rightarrow d_1(0,9)$ in the second.

COMPLEX REQUEST

A complex request is a request that has words such as "and", "or", "not" and are corresponding to fuzzy sets operations – "conjunction", "disjunction" and "complement".

They can be summarized in:

AND	$\wedge$ (conjunction)	min
OR	$\vee$ (disjunction)	max
NOT	$\bar{k}_i$ (complement)	1 - k_i

By representing the searched word in this way we can find the fuzzy set **DS** which contains "all documents with the keyword **ks**" and where the level of adequacy is set to α_{ks}.

For example:

If the request is $ks = k_1 \vee k_2 \wedge \bar{k}_3$ then we can symbolize it as **min[max(k_1,k_2),1-k_3]**. Consequently $\mu_D(ks)=\min[\max(\mu_D(k_1),\mu_D(k_2)),1-\mu_D(k_3)]$ and the fuzzy set **DS** is:

DS=

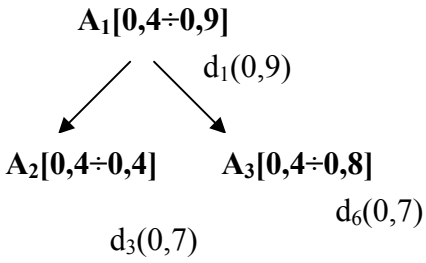
d_1	**d_2**	**d_3**	**d_4**	**d_5**	**d_6**
0,9	**0,3**	**0,4**	**0**	**0,3**	**0,8**

If we set a level α_{ks}**=0,4** then we will find out that the documents **d_1**, **d_3** и **d_6** cover this condition. When we find the max-min and min-max composition of **R** и **DS** we have:

T_1	**ks**
A_1	0,9
A_2	0,4
A_3	0,8

T_2	**ks**
A_1	0
A_2	0
A_3	0

And the result will be:

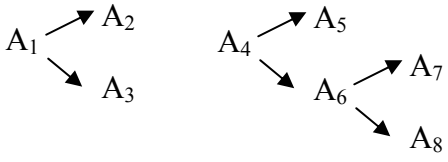


GENERALE CASE IN REAL SYSTEM

The general case is possible as one keyword describes documents of more than one domain and their sub domains, and / or as one document has been classified in more than one domain / sub domain. Consequently when is searched for such a keyword, the result will be two or more different hierarchies grouping the result thematically in domains and their sub domains.

For example:

We have the following hierarchies:



The documents where they consist of are described with binary relation **R=UxD**, where **U** is universe of discourse containing all domains and sub domains we need and **D** is a set of all documents classified in some of sets $\mathbf{A_1,A_2...A_8}$ (in this case).

The relation **R** is presented as:

R	$\mathbf{d_1}$	$\mathbf{d_2}$	$\mathbf{d_3}$	$\mathbf{d_4}$	$\mathbf{d_5}$	$\mathbf{d_6}$	$\mathbf{d_7}$	$\mathbf{d_8}$	$\mathbf{d_9}$	$\mathbf{d_{10}}$	$\mathbf{d_{11}}$	$\mathbf{d_{12}}$	$\mathbf{d_{13}}$	$\mathbf{d_{14}}$
$\mathbf{A_1}$	1	1	0	0	0	0	0	0	0	0	1	0	0	0
$\mathbf{A_2}$	0	0	0	1	0	0	0	0	0	0	0	0	0	0
$\mathbf{A_3}$	0	0	0	0	1	0	0	1	0	0	0	0	1	0
$\mathbf{A_4}$	0	0	0	0	0	1	1	0	0	0	0	0	0	0
$\mathbf{A_5}$	0	0	0	1	0	0	0	0	1	0	0	0	0	0
$\mathbf{A_6}$	1	0	0	0	0	0	0	0	0	0	0	0	0	1
$\mathbf{A_7}$	0	0	1	0	0	0	0	0	0	0	0	1	0	0
$\mathbf{A_8}$	0	0	0	0	0	1	0	0	0	1	0	0	0	0

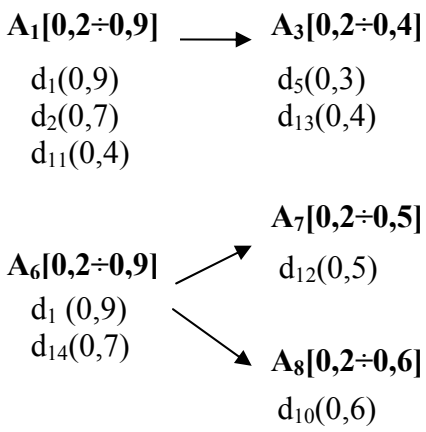
The fuzzy set **K** is a set of all keywords describing the documents. To keep it simple, we will work only with one keyword - κ_1 and its membership function to **D**, which is shown in the binary relation **S**. Now we can find the binary relations $\mathbf{T_1}$ and $\mathbf{T_2}$ that are presenting the optimistic and pessimistic expectations of how κ_1 is describing the following sets $\mathbf{A_1,A_2...A_8}$.

S	$\mathbf{k_1}$
$\mathbf{d_1}$	0,9
$\mathbf{d_2}$	0,7
$\mathbf{d_3}$	0
$\mathbf{d_4}$	0
$\mathbf{d_5}$	0,3
$\mathbf{d_6}$	0
$\mathbf{d_7}$	0
$\mathbf{d_8}$	0
$\mathbf{d_9}$	0
$\mathbf{d_{10}}$	0,6
$\mathbf{d_{11}}$	0,4
$\mathbf{d_{12}}$	0,5
$\mathbf{d_{13}}$	0,4
$\mathbf{d_{14}}$	0,7

T_1	**ks**
$\mathbf{A_1}$	0,9
$\mathbf{A_2}$	0
$\mathbf{A_3}$	0,4
$\mathbf{A_4}$	0
$\mathbf{A_5}$	0
$\mathbf{A_6}$	0,9
$\mathbf{A_7}$	0,5
$\mathbf{A_8}$	0,6

T_2	**ks**
$\mathbf{A_1}$	0
$\mathbf{A_2}$	0
$\mathbf{A_3}$	0
$\mathbf{A_4}$	0
$\mathbf{A_5}$	0
$\mathbf{A_6}$	0
$\mathbf{A_7}$	0
$\mathbf{A_8}$	0

Let **ks= $\mathbf{k_1}$** and **α_{ks}=0,2**. Then the result will be:



This result in a real system would look like a hierarchy of hyperlinks to adjacent list of documents, for example:

Domain1 [0,2÷0,9]
    Sub-domain3 [0,2÷0,5]
Domain4
    Sub-domain6 [0,2÷0,9]
        Sub-domain7 [0,2÷0,5]
        Sub-domain8 [0,2÷0,6]

In this case "*Domain4*" is visualized because it is giving the information about subsequent sub domains.

If we follow the link "*Sub-domain3 [0,2÷0,5]*" for example, the result will be:

Sub-domain3:

Document5 (0,3)
keyword1 (0,3); keyword# (#,#); keyword# (#,#);…
Document13 (0,4)
keyword1 (0,4); keyword# (#,#); keyword# (#,#);…

Where "*Document13 (0,4)*" is hyperlink pointing the real document.

CONCLUSIONS

Knowledge representation using the fuzzy sets theory gives the opportunity to define in a more adequate and clear way the characteristics of an object and classes where it consists of. The presented model gives the possibility to present the result that is classified and helps the user to find the needed domains where he is interested of, grouping the results and in this case filtering out the one's that are not in the required domain. It is given the possibility to set the level of adequacy of the given results.

REFERENCES

[1] Бърнев П., П. Станчев. Размити множества. Народна просвета, София, 1987. 110с.

[2] Гаврилова Т. А., В. Ф. Хорошевский. Базы знаний интеллектуальных систем. Питер, Санкт-Петербург, 2000. 384с.

[3] Дюбуа Д., А. Прад. Теория возможностей – приложения к представлению знаний в информатике. Радио и связь, Москва, 1990. 288с.

[4] Леоненков А.. Нечеткое моделирование с среде METLAB и fuzzyTECH. БХВ-Петербург, Санкт-Петербург, 2003. 736с.

ABOUT THE AUTHORS

Eng. Hristina Moneva, PhD student, Department of Computer Systems and Technology, University of Veliko Turnovo, Phone: + 359 88 9513288, E-mail: xmoneva@ieee.org