medical databases, XML applications,
fuzzy sets ,linguistic summaries

Magdalena BAKALARZ,[*]
Joanna OCHELSKA-MIERZEJEWSKA,[**]
Piotr S. SZCZEPANIAK[*, **]

MEDICAL DATABASES ORGANISATION IN XML LINGUISTIC ENVIRONMENT IMPLEMENTATION

Medical objects are based on different formats, where graphics are the most representative. They provide us with tremendous volumes of the data, very troublesome in management processes. For simplifying these processes, many works for simple and fast descriptors finding were undertaken. Fashionable platform for solving these problems one can be found in XML environment. This technology enables ordering and simple organisation of the database, with friendly linguistic description of the discussed items.

1. INTRODUCTION

Databases' content, given in a variety of formats, as: numerical data, textual files, images, various signal records, provides the user with question of this large data set organisation and its handling. In many circumstances the information is available only if it can be explored by natural language descriptors. Coping with this problem, the transformation of the database content into XML formats has to be considered. It allows us to make a very specific order simplifying the database exploration which becomes independent from an operating system.

Computing machines are very effective while dealing with numbers. For databases containing textual records the exploration process becomes more complex because of its classification and comparison processes concerning natural language expressions. In recent years, the fuzzy sets theory introduced by Zadeh [10] has been proposed as an instrument for effective dealing with this problem. However, for the considered domain-restricted and standardised documents, with respect to their structure and form (of the information entities / records) a recognisable simplification of the task is available.

[*] Medical University, Division of Medical Informatics and Statistics pl. Hallera 1, 90-647 Łódź, Poland
[**] Technical University of Lodz, Institute of Computer Science, ul. Wólczańska 215, 93-005 Łódź , Poland

2. EXTENSIBLE MARKUP LANGUAGE (XML)

XML (*eXtensible Markup Language)* enables us to save data about real or abstractive objects in a formalised and standardised way. It reflects also hierarchical relations between objects. However, the most important advantage of this language is its independence from the platform. The data in XML is possible to process on every machine equipped with any kind of word processor.

The history of XML started in the sixties of the 20th century when IBM Company created the GML (*General Markaup Language*) – the first mark-up language. It made it possible to use all the paper documents needed. At the same time, the GCA Company (*Graphics Communications Associations*) created the GenCode language, which allows ordering documents by studying papers' content. From the eighties, the companies joined their efforts and started creating a standard for defining and using mark-up in textual documents. In 1986 *International Organization for Standardization* published the SGML (*Standard Generalized Markup Language*). From that time, scientists have been working and improving that standard. In 1996, *XML Working Group* presented preliminary version of the simplified SGML. However, the first official version of XML is dated 10th February 1998.

The XML Language in version 1.0 is recommended by W3C (*The World Wide Web Consortium*) consortium. Below, some basic rules and elements that create that standard only used in this system are presented.

1. Mark-up – the element that represents of one data type. It is presented in one of the following forms: <markup> or </markup>. The first one is the opening mark and the second – the closing one. The second type of mark-up is presented by <markup />, where this mark means both opening and closing mark-up. The XML specification distinguish a small from capitals letters. That is why marks like <markup>, <Markup> or <MarkUp> are not found as identical.
2. Attribute – a part of the mark-up, which helps describing a feature of an element in documents. For example, in *<car colour="white">* the attribute is *colour* and the value of the attribute is *white*.
3. The objects must be put into hierarchy as in example 1.

Example 1:

```
<hospital>
        < operating_theatre >
                <surgical_tools>
                </ surgical_tools >
        </ operating_theatre>
        < postoperative_room />
</hospital>
```

There is one important rule – the XML requires closing the second-hand mark before the superior one, like this:

```
<superior>
        <second_hand>
        </second_hand>
```

<superior>

3. LINGUISTIC SUMMARY

The medical databases involve the information description formulated in a natural language. There exists different construction of the sentence, which mean exactly or almost the same. Linguistic summary gives an answer to the question of the relation among the considered data. The answer will be obtained in natural language format. The summary can appear in the form elaborated by Yager [8], concerning numerical data.

Definition 1: The linguistic summary is presented in form

Q P are (have) *S* [*T*]

or, in an extended version

Q of objects being P are S (have property S)
[and correctness of this is of degree T].

The symbols are interpreted in the following manner:
Q – quantity of agreement;
P – subject of the summary;
S – summariser;
T – quality of the summary (measure of validity).

Q is a linguistic quantifier like: "many", "almost all", "few", etc. that shows number of objects defined property. The subject of the summary *P* is determined as the object whose features are described by values from the considered record. The quality of the summary *T* is a real number from the range [0,1] and it is interpreted as a level of truth (confidence) for the given summary.

The content of records in database is also given in natural language. The use of fuzzy sets allows solving the problem within the decision, which membership function is the most suitable for given characteristic feature.

Example 2: *Many patients in hospital are young* [0.88].
Here:
many – *Q*,
patients – *P*,
young – *S*,
[0.88] – *T*.

The way of finding the value of *T* was introduced in the example bellow.

Example 3: The temperature of six patients is: $F = \{f_i\} = \{38.5; 39.1; 39.0; 37.8; 39.6; 38.8\}$. Let the summarisation be performed as the answer to the query:

How many patients have fever ?

Here we have: *S* – *fever*, and *P* – *patient*.

Now, the amount determination *Q*, the quality of the summary *T*, and the suitable fuzzy sets for this linguistic variable *T*, are looked for. First, the membership function, describing the temperature of a patient is defined. This measure is applied to each record. In the next stage, all values of the membership function are counted and then divided by the number of records. Consequently, the average value *r* is obtained. Then, the linguistic variable is defined. The values of *Q*, and *T* are determined on the basis of this definition and the value of *r*.

If the membership functions describing the temperature of the patient is defined as

$$_{fever}\left(f\right)=\begin{cases} 0 & f<37.3 \\ \dfrac{10f-373}{7} & for \quad 37.3\le f<38 \\ 1 & f\ge 38 \end{cases}$$

follows

then the following products will be obtained

The sum of these values is equal to

$Y = 1.0 + 1.0 + 1.0 + 0.71 + 1.0 + 1.0 = 5.71$

By the average value

$$r=\frac{\mathrm{Y}}{\text{number of all patients}}=\frac{5.71}{6}=0.95$$

The definition of the linguistic variable values X = {few, many, almost all} with membership functions as it is shown in Fig.1, the following summarisation results are

$$\mu_{fever}\left(f_1\right)=1.0 \quad \mu_{fever}\left(f_2\right)=1.0 \quad \mu_{fever}\left(f_3\right)=1.0$$
$$\mu_{fever}\left(f_4\right)=0.71 \quad \mu_{fever}\left(f_5\right)=1.0 \quad \mu_{fever}\left(f_6\right)=1.0$$

obtained: almost all patients have fever [0.95].

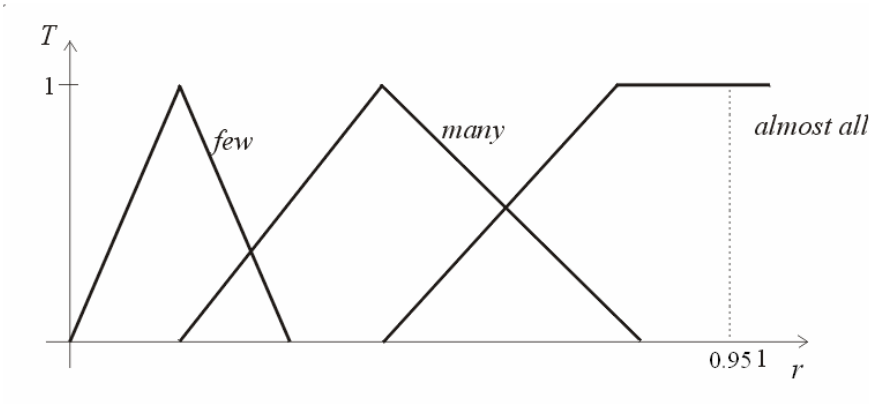

Fig.1. Linguistic variable *X*.

The method can also be applied in more complicated states. For example, in [9], Yager's method is extended to the state where two properties *R* and *S* need specific consideration:

Q P are (have) *R* and *S* [*T*],

and

Q R P are (have) *S* [*T*].

Example 4: *Very few patients are young and in good condition* [0.97],
Very few young patients are in good condition [0.97].

Here:

very few – Q,
patients – P,
young – R,
in good condition – S,
[0.97] – *T.*

Performing summarisation, with respect to some query, requires a proper definition of linguistic variables, determining *Q*, *R* and *S*, as well as computation of *T*.

4. FUZZY SIMILARITY MEASURE FOR WORDS AND SENTENCES

In this section, the fuzzy similarity measures applied to words and sentences are presented. The hypothetical measure was proposed in works [2] and [3, 4, 5]. The measure is an extension of the *n*-gram method, defined in the work [1].

Let S be the set of all words from universe of discourse.

The relation between elements w_1, w_2 of *W* can be determined by the use of membership function $\mu_{RW} : W \times W \to [0,1]$ given by formula

$$\mu_{RW}(w_1, w_2) = \frac{2}{N^2 + N} \cdot \sum_{i=1}^{N(w_1)} \sum_{j=1}^{N(w_1)-i+1} h(i,j) \qquad (1)$$

where

$h(i,j)=1$ if a sub-sequence containing *i* letters of word w_1 and beginning from its *j*-th position in w_1 appears at least once in word w_2; otherwise $h(i,j)=0$;

$N(w_1)$, $N(w_2)$ – number of letters in words s_1, s_1 respectively;

$N=max\{ N(w_1), N(w_2)\}$ – maximum of $N(w_1)$ and $N(w_2)$;

$0.5(N^2 + N)$ – maximal number of the considered sub-sequences.

For asymmetry killing, in relation (1), we define expression $\mu^{\circ}_{RW} : W \times W \to [0,1]$ as follows:

$$\mu^{\circ}_{RW} = min\{\mu_{RW}(w_1, w_2), \mu_{RW}(w_2, w_1)\} \qquad (2)$$

Example 5: The task of solving the comparison of two words, namely w_1 = 'FEVER' and w_2= 'FEVERISH', which is establishing the similarities between them.

Subsequence consisting of steadily increasing characters number, i.e. of length equal to one, then two, and so on (taken from the first word), is built. In the next step, these subsequence is searched for in second word. In case they are found, according to the formula (2), the function $h(i, j)$ is equal to one, otherwise – the function is zero. The possible subsequences are examined like this. All values of function *h*(*i, j*) are added, and the sum is divided by the number of all subsequence characters that have been found.

In our example, in word w_1, there are five single characters in this subsequence:

- w_1 (F, E, V, E, R).

Two-characters strings were introduced like below:

- w_1 (FE, EV, VE, ER).

The further searching results are shown in the following sequences:

- three three-characters sequences of w_1 (FEV, EVE, VER),
- two four-characters sequences of w_1 (FEVE, EVER),
- one five-characters sequence of w_1 (FEVER).

Let us establish the maximum dimention of the characters number in these two words. This value is equal to $N = \max\{ N(FEVER), N(FEVERISH) \} = \max \{ 5, 8 \} = 8$.

These sequences, taken from the first word, are now looked for in the second word w_2, which was described below.

$$\mu_{RW}(FEVER, FEVERISH) = \frac{2}{(N^2+N)} \cdot \sum_{i=1}^{N(w_1)} \sum_{j=1}^{N(w_1)-i+1} h(i,j) =$$

$$= \frac{2}{(8^2+8)} \cdot h(1,1) + h(2,1) + h(3,1) + h(4,1) + h(1,2) + h(2,2) + h(3,2) + h(1,3) + h(2,3) + h(1,4) =$$

$$= \frac{2}{56} \cdot (1+1+1+1+1+1+1+1+1+1) = \frac{20}{56} = 0{,}36$$

Thus, the similarity of the two considered words is 0.36.

For comparison of sequence of words (e.g. sequences) we can compute

$$\mu_{RS}(s_1, s_2) = \frac{1}{N} \sum_{i=1}^{N(s_1)} max_{j=1,\ldots,N(s_2)}\, \mu_R^{\circ}(w_i, w_j) \qquad (3)$$

where: $\mu_{RS} : S \times S \to [0,1]$;

Z – set of all admissible word sequences;

w_i - the word of number *i* in sentence s_1;

w_j - the word of number *j* in sentence s_2;

$\mu_R^{\circ}(w_i, w_j)$ - the value of function μ_{RW} or for the pair (w_i, w_j) given as $\mu_{RW}(w_1, w_2)$ respectively;

$N(s_1), N(s_2)$ - the number of words in sentences s_1, s_2, respectively;

$N = max\{N(s_1), N(s_2)\}$ - the number of words in the longer of the two sentences under comparison. Note, that in terms of the fuzzy sets theory, the μ_{RW} can be interpretedas the membership function.

Symmetry to (3) $\mu_{RS}^{\circ} : S \times S \to [0,1]$ was given in following way:

$$\mu_{RS}^{\circ} = min\{\mu_{RS}(s_1, s_2), \mu_{RS}(s_2, s_1)\} \qquad (4)$$

5. ILLUSTRATIVE IMPLEMENTATION

The introductory stage is the transformation process of the database description into XML items. The authors implemented the MS Access database, although one can use any other databases. Let us consider a list of ten patients with their ID (identification number), AGE (age of the patients) and the disease description in natural language, as in table 1.

Table 1. The database fragment

ID	AGE	HEALT CONDITION
1.	17	Condition of patient - quite good
2.	22	Condition of patient - sufficiently good.
3.	53	Condition of patient - rather good.
4.	24	Condition of patient - fairly good.
5.	29	General condition good.
6.	31	Patient in good condition.
7.	60	General condition of patient - good.
8.	27	General condition of patient - rather good.
9.	19	General condition of patient - very serious.
10.	31	General condition good.

At a first step, each patient's data record was divided into sentences. Next, each sentence is analysed separately, in the XML format. The mark-ups of XML document are the features of the records from database (the attribute, if data is given in numerical form and in the subject of sentence, if data is written in natural language). The databases are often disordered with description given in different grammatical forms and in various formats (i.e. structured or free text). Creating the XML database, the data is ordered and, moreover, the description is transferred into the formalised type (Fig. 2).

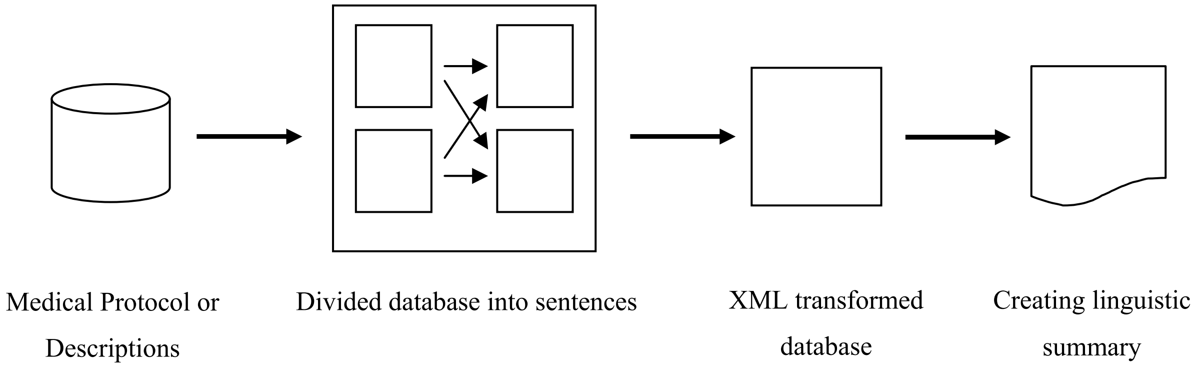
Medical Protocol or Descriptions
Divided database into sentences
XML transformed database
Creating linguistic summary

Fig.2. The multi-step transformation process for XML document

The XML format of the records given in Table 1 was listed in the example of Fig. 3. After that starts the algorithm of the linguistic summary.

The final answer is provided on the basis of a generated summary in final stage. First, however, a linguistic variable and relation between the values calculated as sums of membership functions obtained at the previous stage and the linguistic variable need to be defined. A linguistic summary obtained by means of the described procedure is the answer being searched.

```
<database>
       <record1>
             <age> 17 </age>
             <condition> quite good </condition>
       </record1>
       <record2>
             <age> 22 </age>
             <condition> sufficiently good</condition>
       </record2>
       <record3>
             <age> 53 </age>
             <condition> rather good</condition>
       </record3>
       <record4>
             <age> 24 </age>
             <condition> fairly good</condition>
       </record4>
       <record5>
             <age> 29 </age>
             <condition> good</condition>
       </record5>
       <record6>
             <age> 31 </age>
             <condition> good</condition>
       </record6>
       <record7>
             <age> 60 </age>
             <condition> good</condition>
       </record7>
       <record8>
             <age> 27 </age>
             <condition> rather good</condition>
       </record8>
       <record9>
             <age> 19 </age>
             <condition> very serious</condition>
       </record9>
       <record10>
             <age> 31 </age>
             <condition> good</condition>
       </record10>
</database>
```

Fig.3. The example XML document

From the database given in Table 1 diverse information can be extracted, i.e. answers to diverse queries can be obtained.

Example:

Query: How many of young patients are in good condition?

Note that each information entity related to a particular patient consists of records of both numerical and textual character. Data of numerical type are generally easy to compare and the way of Yagers' summarisation is well-known. When dealing with textual records one needs to use a method, which makes it possible to compare them, more precisely – to compute their similarity degrees. This statement is in accordance with human intuition in which two sentences may be similar to some extent. Qualitative comparison of sentences is based on similarity of words presented in section 4.

The final answer could be like: Many young patient are in good condition [0.4]. More information about creation linguistic summaries can be found in many works [4 -7].

6. CONCLUSIONS

These days, medical databases are considered huge, understandable only for specialists; physicians. That is the reason why modelling this bases is a task that seems to be more and more needed for both, engineers and physicians.

To simplify this problem the authors introduced some investigations of the medical description in the XML formats. At the beginning, one has to separate the patients' description into sentences. Then, the subject and its object, which become adequately, feature and feature's values have to be chosen. After that, it is possible to use linguistic summaries to just describe medical databases in language that is easier to understand for the user. Apart from this, we obtain the database structure, which is running on every operating platform.

BIBLIOGRAPHY

[1] BANDEMER H., GOTTWALD S., Fuzzy Sets, Fuzzy Logic, Fuzzy Methods with Applications. John Wiley & Sons, 1995.

[2] NIEWIADOMSKI A., Appliance of Fuzzy Relations for Text Documents Comparing. Proceedings of the 5th Conference on Neural Networks and Soft Computing, Zakopane, Poland, 2000, 347-352.

[3] NIEWIADOMSKI A., KRYGER P., SZCZEPANIAK P.S., Fuzzy comparison of strings in FAQ answering. In: W. Abramowicz (Ed.), Proceedings of the 7th Business Information Systems, 2004, 355-362.

[4] OCHELSKA J., NIEWIADOMSKI A., SZCZEPANIAK P.S., Linguistic Summaries Applied To Medical Textual Databases. In Journal of Medical Informatics and Technologies, vol. 2, Dept. of Electronics & Computer Systems University of Silesia, MI-117 – MI-124, Ustroń, 2001, ISSN 1642-6037.

[5] OCHELSKA J., SZCZEPANIAK P.S., NIEWIADOMSKI A., Automatic Summarization of Standarized Textual Databases Interpreted in Terms of Intuitionistic Fuzzy Sets. In Grzegorzewski P., Krawczak M., Zadrożny S., Soft Computing Tools, Techniques and Application, Akademicka Oficyna Wydawnicza EXIT, 203-216, Warszawa, 2004, ISBN 83-87674-70-2.

[6] NIEWIADOMSKI A., OCHELSKA J. and SZCZEPANIAK P. S., Interval-Valued Linguistic Summaries of Databases, Control & Cybernetics, vol. 2, 2006, 415-444

[7] SZCZEPANIAK P.S., OCHELSKA J., Linguistic Summaries of Standardized Documents. In Last M., Szczepaniak P.S., Volkovivh Z., Kandel A., Advances In Web Intelligence and Data Mining, Studies in Computational Intelligence, vol. 23, 221-232, Springer-Verlag, Berlin Heidelberg, 2006.

[8] YAGER R.R., Linguistic Summaries as a Tool for Databases Discovery. Workshop on Fuzzy Databases System and Information Retrieval. Yokohama, Japan, 1995.

[9] YAGER R.R., Linguistic Summaries if Data, 3-rd Int. Conference on Information Processing and Management of Uncertainty in Knowledge-Based System, Paris, France, 1990.

[10] ZADEH L.A., Fuzzy Sets, Information and Control, 8, 1965, 338-353.