Derivation of Complex Linguistic Summaries from Databases for a More Human Consistent Descriptive Data Analysis

Gerda Bortsova, Pakizar Shamoi
Department of Computer Science,
Kazakh-British Technical University
gerdabortsova@gmail.com, pakita.shamoi@gmail.com

Abstract

Words provide a natural way of perceiving and manipulating information by humans. Based on that premise, a concept of linguistic description of phenomena arose, and emerged into a large and growing field. In this paper, we present the approach for derivation of meaningful linguistic summaries from databases. We make an emphasis on description of relationships between attributes of a dataset, which can find applications in a number of domains. To demonstrate that, we took a dataset obtained during a research of cervical osteochondrosis among miners and developed a prototype system, which can produce important and interesting conclusions from data, such as "Muscle strength of miners does not change significantly with an increase of work experience from about 10 years to around 15. However, when experience exceeds 20 years, it decreases dramatically."

Introduction

An extensiveness of the use of information technologies in almost all areas of human activity has grown drastically in the past few decades, enabling a collection of huge amounts of various data, which have to undergo some kind of analysis for the benefit of individuals, organizations, or nations. There are many types of data analyses, such as descriptive, exploratory, inferential, predictive, causal, mechanistic, with all of them having different goals and approaches. In this paper we will concentrate our attention on the most frequently used one, that is, descriptive data analysis.

Descriptive statistics and visualization are commonly used methods in the descriptive analysis. However, they constitute only a first step in the process of analysis: the remaining step is interpreting the statistical figures and graphs obtained during the first part. Usually, the interpretation is given in the form of a summary in natural language, such as "men are exposed to alcohol dependence much more frequently, than women". Human-generated summaries often contain subjective and fuzzy terms, such as "much more", "significantly", "large", "young", etc., which eases perception of this information. For example, a sentence "last week, 15,000 units of the product were sold" contains less information for an analyst and sounds less natural compared to "last week, the sales were quite weak".

Based on the fact that natural language is easier to perceive for humans than numbers, a concept of linguistic description, or summarization, of data arose. One of the good examples of work in this area is [6-7], which presents a method that employs fuzzy logic to discover relevant, nontrivial dependencies in multivariate datasets in the form of short sentences that signify some patterns in the data. We believe that this approach, as well as those we mention further in the paper, can be enhanced so as to rich a higher degree of efficacy and widen the area of application of them. Particularly, we argue that such characteristics of a system for linguistic summarization as ability to identify how a variable of interest behaves in relation to the other, or to compare groups of objects in the dataset with respect to a chosen parameter would be very helpful for many applications. Specifically, we believe they are crucial when applied to a field of medical research.

In the next section of this paper we outline the reasons for choosing this research topic. In the part called "Overview of Existing Approaches" we familiarize the reader with modern developments in the field of linguistic summaries derivation and explain novelty of our work. Then we describe our methodology in detail, after which we present results of its implementation in "Application" section and make concluding remarks.

Motivation

In this paper we demonstrate a method to derive important and interesting conclusions from data, which were obtained during a research of cervical osteochondrosis among miners. We examined goals of the original study [1], and found that a system for analysis of data collected by the author should have the following capacity to be truly helpful:

1. Give an easily perceivable and accurate description of a desired piece of data, i.e. of a collection of records in the database, selected by applying certain constraints.
2. Be able to identify differences between distinct subsets of a whole set of records, which may or may not exhibit dissimilar qualities, and provide this comparative analysis in a comprehensive and convenient way.
3. Depict a relation between attributes of a dataset in sufficiently laconic, but condensed form.

We believe that linguistic summarization methods can offer a lot to this field of application and can be used effectively to complement conservative data analytics tools, such as statistics and visualization. However, a methodology we have developed is quite universal and can be applied in any situation, where foregoing functionality is needed.

Overview of Existing Approaches

Emergence of Zadeh's fuzzy sets and logic theory made research towards linguistic description of phenomena possible, with many researchers turning their attention to this field [4, 9, 13].

One of the commonly used methods of linguistic summarization is Yager's approach [10]. Suppose we have:

• V is a quality (attribute) of interest, e.g., age
• $Y = \{y_1, \dots, y_n\}$ is a set of objects (records) that manifest quality V, e.g., a set of employees; hence $V(y_i)$ are values of quality V for object y_i
• $D = \{V(y_1), \dots, V(y_n)\}$ is a set of data ("database")

A summary of a data set consists of:

• A summarizer S (e.g., *young*)
• A quantity in agreement Q (e.g., *most*)
• Truth (validity) T – e.g., 0.7

as, e.g., T(*most* of employees are *young*) = 0.7.

In addition, a set of y's to be described may be restricted by a set of constraints, called filter. Taking into an account all of the above, typical summary looks like:

$Q\ F\ y$'s are S,

e.g., "most (Q) of single (F) employees are young (S)".

Truth value of a summary may be determined in several ways; the basic way is Zadeh's calculus of linguistically quantified propositions [11]:

$$T(Q\ F\ y\text{'s are } S) = \mu_Q[\sum_{i=1}^{n}(\mu_F(y_i) \wedge \mu_S(y_i))/\sum_{i=1}^{n}\mu_F(y_i)],$$

where $\mu_Q(x)$, $\mu_F(x)$ and $\mu_S(x)$ are membership functions of the quantifier, filter and summarizer respectively.

This idea was implemented by Kacprzyk and Zadrozny as an add-on to Microsoft Access, called FQUERY [5, 8]. As fully automatic derivation of linguistic summaries from a sufficiently large database would be very time consuming, they proposed an interactive approach to summarization, in which a user should guess some of the summary's parameters.

In the works mentioned above, as well as in other research works concerning the application of linguistic summaries to data mining, e.g., [2, 3], fuzzy summaries are employed to elicit hidden dependencies in data (fuzzy rules), i.e., in a fashion of exploratory data analysis.

In our research work, we focus on using fuzzy summaries for purposes of describing pieces of data, which are particularly interesting to a user. Firstly, this includes finding a best fitting linguistic label (which may be a compound of labels) for a subset of records in a dataset, e.g., "most of *young* employees have *middle* or *high* income".

In addition, we provide an advanced fuzzy comparison for identifying distinction between subsets of data. This is necessary because, while many approaches to linguistic summarization suggest a good way to assign linguistic labels to groups of records in the database (defined by different filters), it appears quite frequently that we have to directly compare two or more collections of objects' parameters and identify how significant is their difference. For instance, the proposed fuzzy summarization mechanism is able to produce summaries like "Muscle strength asymmetry coefficient of the main group is *dramatically greater* than that of the control group."

As a third novelty, we provide a way to depict relations between variables in a dataset. By a relationship we mean how a variable of interest behaves in relation to the other, for example, "Muscle strength of the miners *drops a little* from *roughly between 725 and 775* to *nearly between 675 and 750* with an increase of age from *around 40 or less* to *roughly between 45 and 50*. Then it *falls drastically* till *approximately between 600 and 625* with an increase of age till *about 55 or greater*."

Also, it is of interest to note that linguistic summaries are most frequently used in the area of business (sales database is a popular example). In this paper we will try to examine what benefits this methodology can offer when applied to the field of medical research.

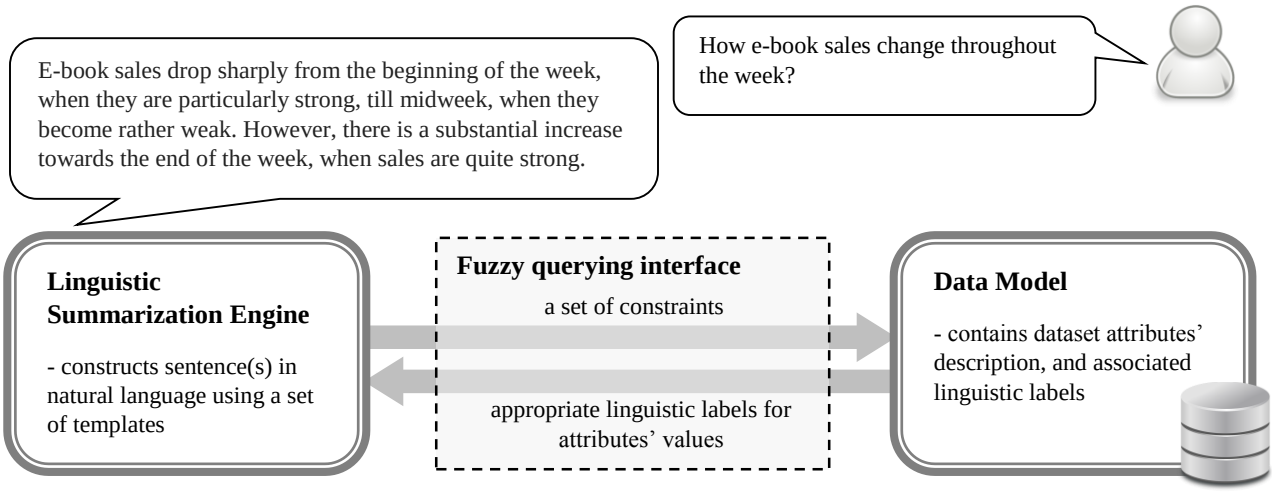

Figure 1. The Main Components of the System.

Methodology

Over the course of our research work, we have developed a system for derivation of linguistic summaries from databases with an emphasis on description of relationships between variables (fuzzy or crisp in nature), or, technically speaking, values of dataset attributes. In particular, the system is able to:

• give an accurate linguistic descriptor for an attribute of subset of data, selected by applying some constraints (fuzzy or exact), for instance, "strong", "roughly between 600 and 650";

• compare groups of objects in a dataset by a chosen parameter, e.g., "this week's sales have *significantly exceeded* that of the previous week";

• discover trends in behavior of a variable in relation to the other one (see example on the Fig. 1).

Fig. 1 reflects main components of the system and their interaction. Data model, which is an abstraction of the database, is responsible for a direct interaction with it. Additionally, it contains a set of linguistic labels (fuzzy partition), associated with each attribute, and their definition (corresponding fuzzy sets). For instance, linguistic variable *sales* has the values: *weak*, *average*, *strong*. Data model is accessed by summarization engine through fuzzy querying interface, which allows obtaining sufficiently good linguistic descriptors (labels) for attributes, manifested by subsets of interest, and difference between these subsets. Summarization engine then uses these pieces of knowledge to build a summary in natural language using built-in sentence structures.

Now let us describe the most important pieces of functionality of our system.

Linguistic Description of a Subset of Data

One of the common tasks in data analysis is to characterize values of an attribute of interest of its subgroups. For instance, a researcher may be interested in values of nerve conduction velocity among stope miners with extensive work experience. Although there are standard procedures for numerical description of data, such as finding mean and standard deviation, we advocate using linguistic labels and approximate numbers and intervals, technically defined by fuzzy sets, as they provide a more human consistent way of information representation.

One of the difficulties we faced in the process of implementation of this feature is concerned with how to define fuzzy partitions for fuzzy variables so as to be able to describe and compare different subsets of the data by a certain parameter. Generally, various groups of objects (selected using different filters) might have highly diverse values of an attribute of interest, or, in contrast, highly narrow variance. On the Fig. 2 you may see a frequency diagram of muscle strength coefficient of asymmetry of miners of a main group (miners of specializations with harmful working conditions, which involve factors that influence development of cervical osteochondrosis) and a control group (miners of specializations with less harmful working conditions). As can be noticed, the main group is "wider" and one would describe it as "approximately in the interval between 1.3 and 1.7", and "roughly between 1.1 and 1.3" for the control group, which is clearly a shorter interval. It is obvious that it is impossible to find an ideal width of a fuzzy interval and a number of fuzzy sets in partition (granularity). Therefore, we either have to define a large number of fuzzy sets which signify approximate intervals of a different length, or find a way to build fuzzy sets on demand, depending on the data itself, but without loosing

their descriptive power, which means they have to be standard in a certain way.

We decided to follow the second path, and designed an algorithm for finding a fuzzy set, combined from predefined fuzzy sets in a partition using disjunction, which fits a subset of data best, with a degree of truth being higher than some threshold value (we took 0.7). In this algorithm we used Zadeh's formula [11] for finding validity of the summary. Conjunction and disjunction ("and" and "or") operators in filter and summarizer parts of the expressions are simple minimum and maximum operators introduced in [12].

Let us demonstrate the results of running this algorithm on some examples beneath. See Fig. 2 and Fig. 3 for comparison (fuzzy intervals identified for the example groups approximately correspond to intervals of high density on the frequency diagrams).

Muscle strength asymmetry coefficient of the main group is approximately between 1.4 and 1.6.

Muscle strength asymmetry coefficient of the control group is about 1.3 or less.

Muscle strength of stope miners is approximately between 675 and 775.

Muscle strength of timbermen is roughly between 650 and 750.

Muscle strength of loader operators is nearly between 625 and 750.

This is important to note that the method described above can be equally efficiently used for crisp and fuzzy subsets of data (selected using fuzzy filter, e.g., "middle-aged" for age).

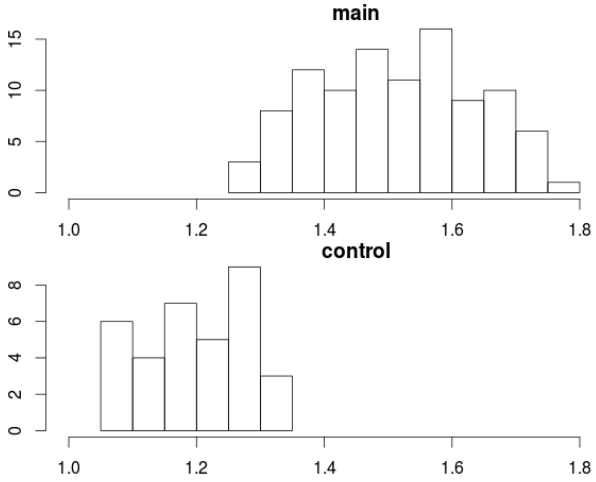

Figure 3. Histograms of Muscle Strength Coefficient of Asymmetry for the Main and Control Group of Miners.

Fuzzy Comparison of Subsets

Another main component of the system is a mechanism of fuzzy comparison, which is used to identify a degree to which two groups of objects are different. Comparison of groups is very useful for many applications. Specifically, in our application it is necessary for such tasks as: determining distinction between the main and control group of workers, as well as between different professions of miners (also with varying labor conditions) to prove negative impact of certain workplace factors on the development of cervical osteochondrosis and overall health, as well as find core influencing factors; compare medical test results before and after treatment, and results of several different therapies.

A standard way to assess difference between two samples is t-test. However, we did not use it for a number of reasons. Firstly, numbers are perceived and interpreted by human worse, than words, and, secondly, there is a problem designing weighted t-test (for cases when subsets are fuzzy).

Instead, we invented our own approach for finding a coefficient of distinction based on linguistic summaries. In brief, we take the best fit fuzzy sets (let's say, S_1 and S_2) corresponding to two subsets of interest (lets call them A and B), and find an average of validity of two summaries: "A is NOT S_2 " and "B is NOT S_1" ("not" is in traditional Zadeh's sense). This method also identifies a direction of this difference (*less* or *greater*).

A linguistic label for a coefficient of distinction is found as a fuzzy set in the partition (defined subjectively) having the largest degree of membership of this coefficient.

Some of the summaries, generated by the program, that implements fuzzy comparison (refer Fig. 2 and 3 for visual comparison) can be seen beneath:

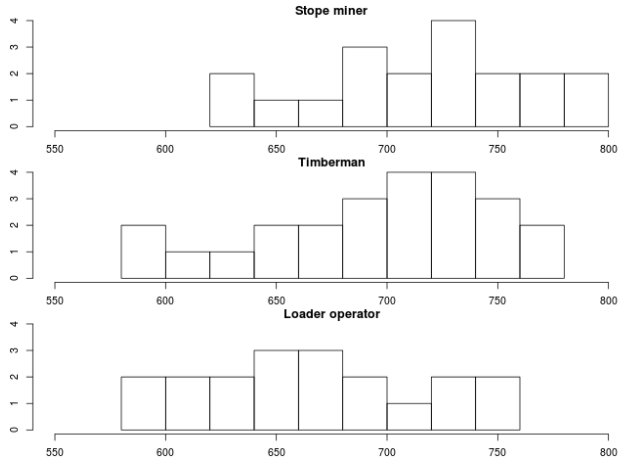

Figure 3. Frequency Diagrams for Muscle Strength of Miners of Different Specializations: Stope Miners, Timbermen and Loader Operators.

Muscle strength asymmetry coefficient of the main group is dramatically greater than that of the control group.

Muscle strength of stope miners is almost identical to that of timbermen.

Muscle strength of loader operators is slightly less than that of stope miners.

Structure of a Linguistic Summary

The simple kinds of summaries given in the previous subsections do not worth much attention per se. However, if combined into more complex structures, they can become a really powerful tool, that enables to depict not only single subgroup or relation between two subgroups' parameters, but a relationship between variables, i.e. attributes of a dataset, which can be fuzzy or crisp in nature. For instance, we are interested in a relationship between age of the workers and muscle strength, which can be represented by a couple of sentences in a natural language (see Fig. 4. for comparison; red dots are means of muscle strength of workers with age intervals loosely corresponding to fuzzy sets in the partition):

Muscle strength of miners drops a little from roughly between 725 and 775 to nearly between 675 and 750 with an increase of age from around 40 or less to roughly between 45 and 50. Then it falls drastically from approximately between 675 and 750 to approximately between 600 and 625 with an increase of age from roughly between 45 and 50 to about 55 or greater.

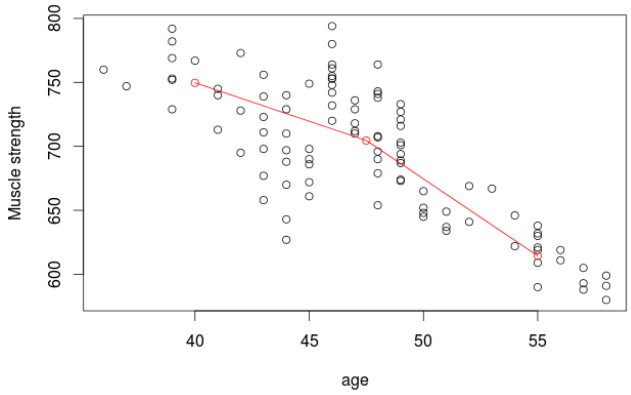

Figure 4. Age vs. Muscle Strength.

Actually, these sentences were generated by a computer program, that compared subsets of data corresponding to all linguistic labels associated with an independent variable (age in the example) by the value of the second parameter (muscle strength), after which it putted the results of each comparison into a simple template:

Y of miners R D from y_1 and y_2 with an increase of X from x_1 to x_2,

where:

- X and Y are an independent and dependent variable respectively, age and muscle strength according to the example;
- x_1 and x_2 are two linguistic labels associated with X domain, that are neighboring if to put all labels in ascending order (*approximately 40 or less* and *between 45 and 50*, *between 45 and 50* and *55 or greater* in the sentence);
- y_1 and y_2 are their respective linguistic descriptors, that were calculated by a method explained earlier in the text (e.g., *roughly between 725 and 775* and *nearly between 675 and 750*);
- R is a relation between subsets y_1 and y_2 (increase, decrease or no change), exemplified by *drop* and *fall* (synonyms) in the above summary;
- D is a linguistic label for a coefficient of distinction (*slightly*, *dramatically*, etc.), found using a method described in the previous subsection.

Adding a Linguistic Flexibility

In order to make our summaries more realistic and human-generated like, we introduced synonyms to the functional parts of a summary and the linguistic labels (refer to Tab. 1). So, for instance, to specify that A *is significantly greater than* B, system may also take any one of its synonyms, such as *substantially*, *noticeably*.

Ingredients	**Synonyms**
Approximate number	about X, around X, roughly greater than X, approximately less than X, nearly between A and B
Direction of difference	
a) a downward trend	decrease, fall, drop, decline
b) an upward trend	increase, rise, climb, grow
Coefficient of distinction	
a) not significant difference	almost the same, identical
b) minor difference	quite significantly, slightly, a little
c) significant difference	significantly, substantially, noticeably
d) major difference	very significantly, dramatically, drastically

Table 1. Summary Ingredients and Synonyms.

Enriching Summary Structure

While the kind of summaries demonstrated above is quite human consistent, it is possible to introduce a number of refinements in the future:

a) Add linking words. Words like *but, furthermore, nevertheless, likewise, again* would allow to increase quality of summaries by highlighting a contrast in trends or their repetition. Examples:

Electromyography test results for the groups of workers with less than 10 years of experience and with 10 -15 years of experience do not differ significantly. However, that for the group of workers with more than 20 years of experience fall sharply from 700-750 to 600-625.

Nerve conduction velocity falls gradually from the age of 35-40 till the age of 40-45. Then it decreases sharply till the age of 45-50 and after that it starts to decrease gradually again.

b) Creating a structure and organization for sentences, summarizing distinctions between groups subsetted using a parameter which is crisp in nature and cannot undergo qualitative comparison, for instance, profession or a method of treatment a worker undergone:

Nerve conduction velocity (NCV) of workers, who have undergone basic therapy treatment, has significantly increased due to the therapy. NCV of workers, who have undergone DENS treatment, has also improved, but to the lesser degree.

Application

In order to enliven the developed methodology and identify its advantages and disadvantages, we implemented some of the aforementioned functionality of the proposed system in Ruby programming language, together with Ruby on Rails web framework, that we utilized for easier database manipulation and user interface design. Let us now demonstrate our results.

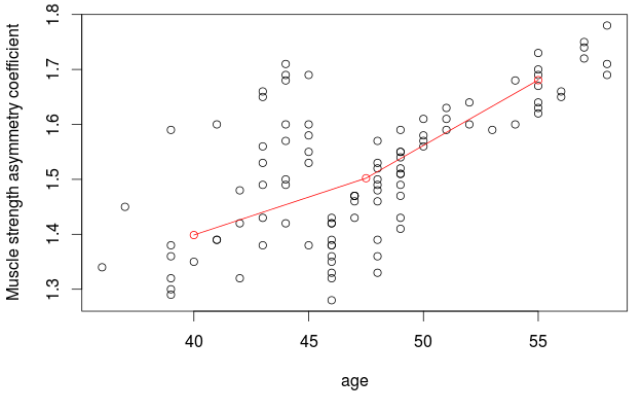

Figure 5. Example #1.

Example 1. Age and muscle strength coefficient of asymmetry (refer to Fig. 5 for comparison).

Muscle strength asymmetry of miners rose quite significantly from approximately 1.5 or less to roughly between 1.4 and 1.6 with an increase of age from approximately 40 or less to approximately between 45 and 50. Then it rose drastically from approximately between 1.4 and 1.6 to roughly 1.6 or greater with an increase of age from roughly between 45 and 50 to roughly 55 or greater.

Example 2. Age and muscle strength coefficient of asymmetry with a finer scale for age (see Fig. 6 for comparison).

Muscle strength asymmetry of miners climbed quite significantly from roughly less than 1.5 to approximately between 1.4 and 1.6 with an increase of age from approximately less than 40 to roughly 45. Then it climbed a little from approximately between 1.4 and 1.6 to roughly between 1.5 and 1.6 with an increase of age from about 45 to roughly 50. Then it climbed very significantly from roughly between 1.5 and 1.6 to approximately greater than 1.6 with an increase of age from around 50 to around 55. Then it was almost the same with an increase of age from roughly 55 to approximately greater than 55.

Earlier in this paper, we mentioned several use cases of our system. Let us provide a more comprehensive list of its potential applications with accordance to the goals of the research [1]. Namely, the system can be used to:

1. Compare health parameters of miners of different specializations with varying labor conditions to determine the most and the least harmful occupations; as well as to compare main and control group (workers, which labor conditions are determined as very heavy and those with not heavy working conditions).

2. Find relationships between various parameters of patients, such as age and work experience, and results of medical tests, so as to prove negative impact of miner's workplace factors on the development of cervical osteochondrosis.

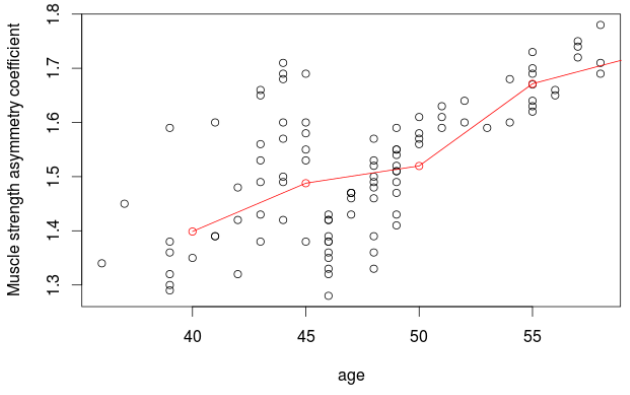

Figure 6. Example #2.

3. Analyze a relationship of potentially unfavorable workplace factors that influence development of cervical osteochondrosis, like heavy physical labor, vibration (coming from drilling machine, loader, etc.), microclimate, noise, and degradation of functional parameters of the body.

Concluding remarks

In this paper we presented a system for derivation of linguistic summaries from databases, which possesses functionality of linguistically describing relationships between attributes of a dataset and distinction in properties of subsets of data. Main challenges we have faced were construction of a good descriptor for a subset, deciding on the metric for significance of difference between subsets, and designing the summary structure. In our future works we plan to enrich summary structure with new templates and components like linking words. Also, we plan to examine a potential of our methodology for application in areas other than medical research, as we believe that it is universal and can be applied in any context where linguistic summaries can aid in decision support or other purposes, involving data analysis.

References

[1] Bortsova, S.R. 2010. Клинико-функциональная оценка компрессионно-корешковых нарушений шейного остеохондроза у горнорабочих и немедикаментозная коррекция. Candidate of Sciences dissertation, Karagandy, Kazakhstan.

[2] Fiot, C. and Laurent, A. and Teisseire, M. and Laurent, B. 2006. Why Fuzzy Sequential Patterns can Help Data Summarization: An Application to the INPI Trade Mark Database. In *Proceedings of the IEEE International Conference on Fuzzy Systems*, 3596-3603. Vancouver, Canada.

[3] Laurent, A., Bouchon-Meunier, B. and Doucet, A. 2001. Towards Fuzzy-OLAP Mining, In *Proceedings of Workshop PKDD 'Database Support for KDD'*, 51-62. Freiburg, Germany.

[4] Liétard, L. 2012. A functional interpretation of linguistic summaries of data. *Information Sciences 188*:1–16.

[5] Kacprzyk, J., Yager, R.R., and Zadrożny, S. 2001. Fuzzy Linguistic Summaries of Databases for an Efficient Business Data Analysis and Decision Support. In *Knowledge discovery for business information systems*, 129–152. Boston, Mass.: Kluwer Academic Publishers.

[6] Kacprzyk, J., Yager, R.R., and Zadrożny, S. 2000. A Fuzzy Logic Based Approach to Linguistic Summaries of Databases. *International Journal of Applied Mathematics and Computer Science 10*:813–834.

[7] Kacprzyk, J., and Yager, R.R. 2001. Linguistic Summaries of Data Using Fuzzy Logic. *International Journal of General Systems 30*:133–154.

[8] Kacprzyk, J., and Zadrożny S. 2001. SQL and FQUERY for Access. In *Proceedings of IFSA/NAFIPS*, 2464-2469. Vancouver, Canada: IEEE.

[9] Pei, Z., Xu, Y., Ruan, D., and Qin, K.. 2009. Extracting complex linguistic data summaries from personnel database via simple linguistic aggregations. *Information Sciences 179(14)*:2325–2332.

[10] Yager, R.R. 1982. A new approach to the summarization of data. *Information Sciences 28*:69–86.

[11] Zadeh, L.A. 1983. A computational approach to fuzzy quantifiers in natural languages. *Computer Mathematics with Applications 9*:149–183.

[12] Zadeh, L.A.. 1965. Fuzzy sets. *Information and Control 8*:338–353.

[13] Zadrożny, S., and Kacprzyk, J. 2011. From a static to dynamic analysis of weblogs via linguistic summaries. In *Proceedings of 2011 IFSA World Congress AFSS International Conference*, 110–119. Surabaya and Bali, Indonesia.