Journal of Theoretical and Applied Computer Science
ISSN 2299-2634 (printed), 2300-5653 (online)

Vol. 8, No. 1, 2014, pp. 15–34
http://www.jtacs.org

On multi-subjectivity in linguistic summarization of relational databases

Adam Niewiadomski, Izabela Superson

Institute of Information Technology, Lodz University of Technology

Adam.Niewiadomski@p.lodz.pl, 143104@edu.p.lodz.pl

Abstract: *We focus on one of the most powerful computing methods for natural-language-driven representation of data, i.e. on Yager's concept of a* linguistic summary of a relational database *(1982). In particular, we introduce an original extension of that concept: new forms of linguistic summaries. The new forms are named* Multi-Subject linguistic summaries*, because they are constructed to handle more than one set of subjects, represented by related sets of records/objects collected in a database, like "cars, bicycles and motorbikes" (within vehicles), "male and female" (within people), e.g.* More boys than girls play football well*. Thanks to that, the generated linguistic summaries – quasi-natural language sentences – are more interesting and human-oriented. Moreover, they can be applied together with the classic forms od summaries, to enrich naturality of comments/descriptions generated. Apart from traditional interpretions linguistic summaries in terms of fuzzy logic, we also introduce some higher-order fuzzy logic methods, to extend possibilities of representing too complex or too ill-defined linguistic terms used in generated messages. The new methods are applied to a computer system that generates natural language description of numeric data, that makes them possible to be clearly presented to an end-user.*

Keywords: *Multi-Subjectivity in relational databases, linguistic summaries of databases, Multi-Subject linguistic summaries, fuzzy sets*

1. Linguistic summaries of relational databases: an overview on ideas and literature

More than thirty years ago, R. R. Yager proposed the idea of *a linguistic summary of a (relational) database* [1], e.g. *More than half of basketball players are very tall*. This simple concept appeared to be a direct answer to people's needs for quick and friendly receiving of large amounts of data and/or information. What is the most important, the idea does not refer to any of terse statistical method for aggregating data (the mean, variation, standard deviation, etc.) but on fuzzy models of natural language expressions. Even if these expressions are less precise than numbers, e.g. *more than half of objects* instead of 55.6% *of objects* or *a very tall boy* instead of 195*cm-tall-boy*, they are commonly understood and provide knowledge on what the summarized data mean.

The concept of a linguistic summary is based on Zadeh's calculus of linguistically quantified propositions (statements) [2]. There are two basic forms of linguistic summaries (based on two forms of linguistically quantified propositions, respectively) presented in the literature [1, 3, 4, 5, 6]:

$$Q\ P \text{ are/have } S\ [T] \tag{1}$$

e.g. *Many boys are tall* $[0.83]$, and

$$Q\ P \text{ being } W \text{ are/have } S\ [T] \tag{2}$$

e.g. *Many boys who are teenagers, are tall* $[0.63]$. In both (1) and (2) forms, Q is a *quantity in agreement*, e.g. *Many, More than 900*, represented by an aggregation operator, e.g. fuzzy quantifier or an OWA operator [7]. P is the subject of the summary, e.g. men, cars, or any other objects described in the summarized database. S is a *summarizer* – a linguistic expression for properties of the objects, represented by a fuzzy set. The W symbol, appearing only in form (2), is *a qualifier*, represented by a fuzzy set, that determines additional and/or specific properties of the objects that the summary deals with. $T \in [0, 1]$ is *a degree of truth* and it determines how good (how informative, how true) the summary is; values of T are evaluated according to the Zadeh calculus of linguistically quantified propositions and/or to another different methods of evaluating [5, 8].

Obviously, we are unable here to present or even mention all methods and applications of linguistic summarization of databases, e.g. [9, 10]. Moreover, we are not able to enumerate all the concepts for data summarization that are based on fuzzy sets but take into account assumptions different than the Yager originals, e.g. [11, 12, 13, 14, 15].

The main scope of the paper is to introduce a concept of **Multi-Subject Linguistic Summary** of a relational database. "Multi-Subject" means that this new form – in comparison to classic forms (1) and (2) – is linked to more than one subject P, for instance to P_1 and P_2 represented by related subsets of records/objects in a database, like "cars, bicycles and motorbikes" (in a database describing vehicles), or "boys and girls" (if a set of data describes young people), e.g. *More boys than girls play football well*. The new forms of summaries do not replace or supersede forms (1) and (2), but are intended to be an interesting extension of methods of generating descriptions of datasets using quasi-natural language. They are supposed to enrich naturality of comments/descriptions generated by computer systems that deal with large amounts of data.

The proposed summarization methods, considering some future work and research, may apply not only to data collected and stored using the relational model. This model is assumed because of a necessary „table view" of a set of data, i.e. we distinguish „rows" (tuples, records) and „columns" (attributes). Besides, the „subject of the summary", P" can be easily and intuitively associated with a tuple/record/row in a table. Such a „layout" of data being summarized is also assumed when presenting linguistic summarization forms, algorithms of evaluating their degrees of truth, and splitting datasets with respect to chosen attributes. Moreover, the convinience of the reader, who is for sure familiar with the relational model of databases, is taken into account. Please note, that in this paper we do not discuss the origin of data or methods of collecting and storing them.

The data may be stored in both relational and non-relational manners. We either do not take into account methods pre-processing of summrized data (e.g. executing SQL queries, analysing paths in graph or hierarchical databases, conversions between different data models or formats, etc.) which are out of soft computing methods we intend to present ere. However, these future directions of research are promising and may lead to elaborating a more general methods of data linguistic summarization, we mean methods independent of a particular data model.

Hence, the rest of the paper is organized as follows: In Section 2, we explain what we mean by "multi-subjectivity" in a database. A simple algorithm for selecting different subjects from a database is presented. These two or more subjects are represented by non-fuzzy sets of

tuples, or can be, if necessary, results of any other process like selecting, querying or filtering tuples with respect to chosen values and/or attributes, e.g. male and female as values of the Gender attribute. The subjects are then characterized by labels (summarizers) represented by fuzzy sets, to enable aggregation and constructing summaries. Next, in Section 3, the concept of *a Multi-Subject Linguistic Summary* of a relational database is introduced. We propose five new forms of linguistic summaries (Sections 3.1-3.5), and each of them is linked to at least two subjects the database collects data on, e.g. to P_1 *and* P_2 or to P_1 *in comparison to* P_2. Evaluation methods for these new forms of summaries are also provided; traditionally, we call them *degrees of truth*, though extending and adopting some other known measures is also possible, cf. [6, 8]). Section 4 presents two experiments, both based on multi-subject linguistic summaries. In the first experiment we are trying to discover how children age and gender is related to their height. The second experiment attempt to discover associations between gender, age, education and monthly income depending on the region. We demonstrate sample output of the application for the chosen databases, and how users may affect on summaries generated by the software. Finally, there are conclusions on the presented methods drawn in Section 5.

2. Multi-subjectivity in relational databases: new possibilities of data linguistic summarization

We refer to the traditional model of relational databases by Codd [16]. We assume that a database consists of tables being sets of tuples (usually called "records"), and one tuple is a representation of one real object (a child, person, car, transaction, etc.). Table $\mathbb{D}$ consists of tuples d_i of the same subject, $i = 1, 2, \ldots, m$, and $m \in \mathbb{N}$ is the number of tuples in $\mathbb{D}$. Each tuple d_i consists of $n \in \mathbb{N}$ values of attributes $V_1, \ldots, V_n$ and the domains of the attributes are $\mathbb{X}_1, \ldots, \mathbb{X}_n$, respectively. The values of attributes express properties of objects, e.g. height, salary, price, date, etc. and they are treated as "columns" of the table. The value of attribute V_j for object y_i, is denoted as $V_j(y_i) \in \mathbb{X}_j$, $i \in \{1, 2, \ldots, m\}$, $j \in \{1, 2, \ldots, n\}$. Hence, the database D collecting information on elements from $\mathbb{Y} = \{y_1, \ldots, y_m\}$ is in the form of:

$$\mathbb{D} = [d_1, d_2, \ldots, d_m]^T \tag{3}$$

where $d_i = \langle V_1(y_i), V_2(y_i), \ldots, V_n(y_i)\rangle$.

Now is important to mention that objects $\{y_1, \ldots, y_m\}$ from set $\mathbb{Y}$ are the subject of a linguistic summary, see (1) and (2). The concept of a multi-subject linguistic summary is based on **possible splitting set $\mathbb{Y}$ into two or more subsets** but elements in these subsets are still described by the same attributes (columns). So it makes possible to make comparisons between subjects on the base of linguistically expressed values of other attributes, e.g. splitting set children into subsets "boys" and "girls" makes it possible to compare height or weight for these two subjects.

The process of "splitting" dataset into two or more subsets representing selected subjects is described as follows:
FOR $i := 1$ TO m

1. Select attribute V_j, $j = 1, 2, \ldots, n$ in $\mathbb{D}$. This attribute determines whether a given object is a member of one of subjects that are to be distinguished.
2. Get object y_i
3. Get $V_j(y_i)$ and add object y_i to the corresponding subset.

Table 1. A sample database $\mathbb{D}$ collecting information on children in school age

ID	Gender	Age	Height
1.	girl	7	130
2.	boy	8	120
3.	boy	13	150
4.	girl	8	140
5.	girl	18	160

Table 2. The part of dataset $\mathbb{D}$ presented in Table 1, filtered for attribute "Gender"="boy"

ID	Gender	Age	Height
2.	boy	8	120
3.	boy	13	150

Table 3. The part of dataset $\mathbb{D}$ presented in Table 1, filtered for attribute "Gender"="girl"

ID	Gender	Age	Height
1.	girl	7	130
4.	girl	8	140
5.	girl	18	160

A sample database $\mathbb{D}$ in the form of (3) is shown by Table 1. It is a part of a larger database, then summarized in the example presented in Section 4. The table illustrates the possibility of extracting two sets of subjects for multi-subject summaries; in this case, its attribute Gender and its two values: "boy" and "girl", that allow us to "split" the set of data into two subsets, exemplified by Table 2 and Table 3, respectively:

It must be underlined that Table 2 and 3 do not represent real database tables stored separately in a database management system; such storage could appear inefficient and non-optimal, especially, with respect to normal forms of relational database tables, a popular optimisation criteria for databases. The presented tables are only results of filtering operations performed on $\mathbb{D}$ (represented by Table 1) with respect to values of a chosen attribute, here: "Gender", for both "boys" and "girls" values.

What is crucial for the main idea of the paper, i.e. for multi-subject linguistic summaries, is that **(at least) two separated sets of objects**, previously stored as one set in $\mathbb{D}$, are now distinguished. These sets represent **different subjects P_1, P_2, ... of multi-subject linguistic summaries** that are now presented in Section 3.

3. New forms of summaries: Multi-Subject Linguistic Summaries

Note that none of the older forms of linguistic summaries, i.e. (1) and (2), is able to represent the relations or associations between different groups of objects and/or their properties, e.g. between boys and girls in relation to their height, age, etc. For those non-multi-subject methods, the only opportunity is to generate summaries that includes the pre-selected set of objects, e.g. boys or girls, as qualifier W, see (2), e.g. *About half of BOYS are tall*, where *BOYS* is a qualifier.

On the other hand, these relations can be easily discovered and expressed in an interesting way using multi-subject linguistic summaries. Five forms of expressions that are linked to more than one subject (in the sense of "subset of objects/records/tuples") are now presented.

The proposed methods of evalutaing degrees of truth are to extend similar methods originating from the Zadeh calculus of linguistically quantified propositions and „classic" forms of linguistic summaries. They are all based on the concept of a cardinality of a „σcount" of fuzzy sets/type-2 fuzzy sets representing summarizers S and qualifiers W in forms of linguistic summaries. Analogously, the cardinalities of sets of subjects of summaries, i.e. number of elements (rows/tuples) representing subjects M_{P_i}, $i = 1, 2, 3....$

3.1. The first form of a multi-subject linguistic summary

The first form of a multi-subject linguistic summary is proposed:

$$Q\ P_1 \textit{ in comparison to } P_2 \textit{ are } S_1[T] \tag{4}$$

where Q is a fuzzy quantifier, P_1 and P_2 are the subjects of the summary and S_1 is a summarizer, represented by a fuzzy set. The degree of truth of summary (4) is evaluated with formula (5):

$$\begin{aligned} &T(Q\ P_1 \textit{ in comparison to } P_2 \textit{ are } S_1) = \\ &= \mu_Q \left(\frac{\frac{1}{M_{P_1}} \Sigma\textit{-count}(S_{1_{P_1}})}{\frac{1}{M_{P_1}} \Sigma\textit{ -count}(S_{1_{P_1}}) + \frac{1}{M_{P_2}} \Sigma\textit{ -count}(S_{1_{P_2}})} \right) \end{aligned} \tag{5}$$

where:

$$\Sigma\textit{-count}(S_{1_{P_1}}) = \sum_{i=1}^{m} \{u_{S_1}(d_i) : d_i \in^* P_1\} \tag{6}$$

and Σ-*count*$(S_{1_{P_2}})$ – analogously. The notation $d_i \in^* P_1$ means that d_i is a tuple representing P_1 subject. M_{P_1} and M_{P_2} are numbers of tuples representing subjects P_1 and P_2, respectively:

$$M_{P_1} = \sum_{i=1}^{m} {t_i}_{P_1} \tag{7}$$

where:

$${t_i}_{P_1} = \begin{cases} 1, & \text{if } d_i \in^* P_1 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

For instance:

$${t_i}_{\text{boys}} = \begin{cases} 1, & \text{if } V_j(d_i) = \text{"boy"} \\ 0, & \text{if } V_j(d_i) = \text{"girls"} \end{cases} \tag{9}$$

and $V_j = Gender$.

For summary which uses type-2[1] fuzzy sets to describe quantifiers, summarizers and qualifiers, degree of truth is given with formula (10):

$$\begin{aligned} &T(\widetilde{Q}\ P_1 \textit{ in comparison to } P_2 \textit{ are } \widetilde{S}_1) = \\ &= \mu_{\widetilde{Q}} \left(\frac{\frac{1}{M_{P_1}} \text{ card } (\widetilde{S}_{1_{P_1}})}{\frac{1}{M_{P_1}} \text{ card } (\widetilde{S}_{1_{P_1}}) + \frac{1}{M_{P_2}} \text{ card } (\widetilde{S}_{1_{P_2}})} \right) \end{aligned} \tag{10}$$

[1] Higher order fuzzy sets in linguistic summarization are described in [17, 18, 19, 20]

Table 4. Database $\mathbb{A}$ collecting information on children in school age

ID	Gender	Height
1.	boy	155
2.	boy	140
3.	boy	165
4.	girl	160
5.	girl	130
6.	girl	135

where:

$$\text{card}\ (\widetilde{S}_{1_{P_1}}) = \sum_{i=1}^{m} \max\{u_{\widetilde{S}_1} : \mu_{\widetilde{S}_1}(d_i, u_{\widetilde{S}_1}) = 1 \wedge d_i \in^* P_1\} \tag{11}$$

and card$(\widetilde{S}_{1_{P_2}})$ – analogously.

An example of a summary in the form of (4) is now given:

$$\textit{Most of boys in comparison to girls are tall}\ [0.56] \tag{12}$$

where Q =*most of*, P_1 =*boys*, P_2 =*girls*, S_1 =*tall*.

Example 3.1. *Example of calculating degree of truth T for summary 12: Lets assume that T for summary12 will be calculated on simplified database $\mathbb{D}$ contaning three tupples representing boys and three represening girls.*

On ground of table $\mathbb{A}$ we have $M_{P_1} = M_{P_2} = 3$ because both subjects of boys and girls are equal and count three tupples. If label tall is represented by fuzzy set with triangular membership function 33, then Σ-count$(S_{1_{P_1}}) = 0.22 + 0 + 0.67 = 0.89$ and Σ-count$(S_{1_{P_2}}) = 0.44 + 0 + 0 = 0.44$. Using equation 5:

$$T(Q\ P_1 \textit{ in comparison to } P_2 \textit{ are } S_1) = \ = \mu_Q \left(\frac{\frac{0.89}{3})}{\frac{0.89}{3}) + \frac{0.44}{3})} \right) = \mu Q(0.67) \tag{13}$$

In the last step we need to calculate $\mu Q(0.67)$ using membership function of quantifier ,,most of" presented on the figure 1. A we can see on the graph, $\mu Q(0.67) \approx 0.56$.

3.2. The second form of a multi-subject linguistic summary

The second form of a multi-subject summary proposed here is given:

$$Q\ P_1 \textit{ in comparison to } P_2 \textit{ being } S_2 \textit{ are } S_1[T] \tag{14}$$

where S_2 is a qualifier, cf. (2). The degree of truth of the summary is evaluated via formula (15):

$$\begin{gathered} T(Q\ P_1 \textit{ in comparison to } P_2 \textit{ being } S_2 \textit{ are } S_1) = \\ = \mu_Q \left(\frac{\frac{\Sigma\textit{-count}(S_{1_{P_1}} \cap S_{2_{P_1}})}{M_{P_1}}}{\frac{\Sigma\textit{-count}(S_{1_{P_1}} \cap S_{2_{P_1}})}{M_{P_1}} + \frac{\Sigma\textit{-count}(S_{1_{P_2}} \cap S_{2_{P_2}})}{M_{P_2}}} \right) \end{gathered} \tag{15}$$

where Q is a relative quantifier, P_1 and P_2 are the subjects of the summary, S_2 is a qualifier related to both P_1 and P_2 subjects, S_1 is a summarizer,

$$\begin{aligned} &\Sigma\text{-}\textit{count}(S_{1_{P_1}} \cap\ S_{2_{P_1}}) = \\ &= \sum_{i=1}^{m} \min\{\mu_{S_1}(d_i), \mu_{S_2}(d_i)\},\ d_i \in^* P_1 \end{aligned} \tag{16}$$

and Σ-*count*$(S_{2_{P_1}})$, Σ-*count*$(S_{2_{P_2}})$, $d_i \in^* P_1$ – analogously to (4).

For type-2 summary degree of truth is evaluated:

$$\begin{aligned} &T(\widetilde{Q}\ P_1 \textit{ in comparison to } P_2 \textit{ being } \widetilde{S}_2 \textit{ are } \widetilde{S}_1) = \\ &= \mu_{\widetilde{Q}}\left(\frac{\frac{1}{M_{P_1}} \text{ card } (\widetilde{S}_{1_{P_1}} \cap \widetilde{S}_{2_{P_1}})}{\frac{1}{M_{P_1}} \text{ card } (\widetilde{S}_{2_{P_1}}) + \frac{1}{M_{P_2}} \text{ card } (\widetilde{S}_{2_{P_2}})}\right) \end{aligned} \tag{17}$$

where $\widetilde{Q}$ is a relative quantifier, P_1 and P_2 are the subjects of the summary, $\widetilde{S}_2$ is a qualifier related to both P_1 and P_2 subjects, $\widetilde{S}_1$ is a summarizer,

$$\begin{aligned} \text{card}(\widetilde{S}_{1_{P_1}} \cap\ \widetilde{S}_{2_{P_1}}) = \sum_{i=1}^{m} \min\Big\{ &\max\{u_{\widetilde{S}_1} : \mu_{\widetilde{S}_1}(d_i, u_{\widetilde{S}_1}) = 1 \wedge d_i \in^* P_1\}, \\ &\max\{u_{\widetilde{S}_2} : \mu_{\widetilde{S}_2}(d_i, u_{\widetilde{S}_2}) = 1 \wedge d_i \in^* P_1\}\Big\} \end{aligned} \tag{18}$$

and card$(\widetilde{S}_{2_{P_1}})$, card$(\widetilde{S}_{2_{P_2}})$, $d_i \in^* P_1$

An example of a summary in the form of (14) is now presented:

$$\textit{About two-third of boys in comparison to girls being teenagers, are tall } [0.390] \tag{19}$$

where Q =*about two-third*, P_1 =*boys*, P_2 =*girls*, S_1 =*tall*, S_2 =*teenagers*.

Summaries in form (14) allow us to retrieve information about selected subjects' features S_1, according to other subjects conditions (specific features that both subjects must posses). It means that in this case, the tuples taken into account represent boys and girls who are qualified by S_2 as teenagers.

3.3. The third form of a multi-subject linguistic summary

The third form of a multi-subject linguistic summary is proposed as:

$$Q\ P_1 \textit{ being } S_2 \textit{ in comparison to } P_2 \textit{ are } S_1[T] \tag{20}$$

and its degree of truth is evaluated with formula (21).

$$\begin{aligned} &T(Q\ P_1 \textit{ being } S_2 \textit{ in comparison to } P_2 \textit{ is } S_1) = \\ &= \mu_{\widetilde{Q}}\left(\frac{\frac{1}{M_{P_1}}\Sigma\text{-}\textit{count}(S_{1_{P_1}} \cap S_{2_{P_1}})}{\frac{1}{M_{P_1}}\Sigma\text{-}\textit{count}(S_{1_{P_1}}) + \frac{1}{M_{P_2}}\Sigma\text{-}\textit{count}(S_{1_{P_2}})}\right) \end{aligned} \tag{21}$$

where Q is a relative quantifier, P_1 and P_2 are the subjects of the summary, S_2 is a qualifier referring only to subject P_1 and S_1 is a summarizer.

The equation for type-2 summaries:

$$T(\widetilde{Q}\ P_1 \textit{ being } \widetilde{S}_2 \textit{ in comparison to } P_2 \textit{ is } \widetilde{S}_1) = \\ = \mu_{\widetilde{Q}}\left(\frac{\frac{1}{M_{P_1}}\ \text{card}\ (\widetilde{S}_{1_{P_1}} \cap \widetilde{S}_{2_{P_1}})}{\frac{1}{M_{P_1}}\text{card}\ (\widetilde{S}_{1_{P_1}}) + \frac{1}{M_{P_2}}\text{card}\ (\widetilde{S}_{1_{P_2}})}\right) \tag{22}$$

where $\widetilde{Q}$ is a relative quantifier, P_1 and P_2 are the subjects of the summary, $\widetilde{S}_2$ is a qualifier referring only to subject P_1 and $\widetilde{S}_1$ is a summarizer. An example of such a summary is given (20):

$$\textit{About half of boys being teenagers in comparison to girls, are tall}\ [0.256] \tag{23}$$

where Q =*about half*, P_1 =*boys*, P_2 =*girls*, S_1 =*tall*, S_2 =*teenagers*.

Summaries in the form of (20) allows users to retrieve information on some selected features of subjects, according to chosen conditions given for subject P_1 only (i.e. some specific features that only subject P_1 must fulfill). It means that tuples taken into account by the summary represent both P_1 and P_2 subjects, i.e. boys and girls, but only P_1 is additionally qualified by S_2 (here: as teenagers).

3.4. The fourth form of a multi-subject linguistic summary

The fourth form of a multi-subject summary is proposed:

$$\textit{More } P_1 \textit{ than } P_2 \textit{ are } S_1[T] \tag{24}$$

This form does not involve any quantifier. The degree of truth of the summary is given by formula (25):

$$T(\textit{More } P_1 \textit{ than } P_2 \textit{ are } S_1) = \frac{\Sigma\textit{-count}(S_{1_{P_1}})}{\Sigma\textit{ -count}(S_{1_{P_1}}) + \Sigma\textit{ -count}(S_{1_{P_2}})} \tag{25}$$

where P_1 and P_2 are the subjects of the summary, M_{P_1} and M_{P_2} are the numbers of tuples representing subjects P_1 and P_2, $d_{i_{P_1}} : d_i \in^* P_1 \wedge d_{i_{P_2}} : d_i \in^* P_2$.

Degree of truth for type-2 summary:

$$T(\textit{More } \ P_1 \textit{ in comparison to } P_2 \textit{ are } \widetilde{S}_1) = \frac{\text{card}\ (\widetilde{S}_{1_{P_1}})}{\text{card}\ (\widetilde{S}_{1_{P_1}}) + \ \text{card}\ (\widetilde{S}_{1_{P_2}})} \tag{26}$$

An example of such a summary is given:

$$\textit{More boys than girls are tall}\ [0.756] \tag{27}$$

where P_1 =*boys*, P_2 =*girls* and S_1 =*tall*.

Summaries in the form of (24) allow users to compare two different subjects without using any additional measures or fuzzy models, e.g. quantifiers. This method is useful for generating simple, quick and very intuitive summaries.

3.5. The fifth form of a multi-subject linguistic summary

The last of presented forms of multi-subject linguistic summaries is based on forms presented above. This shows flexibility of multi-subject forms as it is not limited to two subjects only:

$$Q\ P_{1_1}, P_{1_2}, ..., P_{1_n}\ \textit{in comparison to}\ P_{2_1}, P_{2_2}, ..., P_{2_n}\ \textit{are}\ S_1[T] \tag{28}$$

where $P_{1_1}, P_{1_2}, ..., P_{1_n}$ is the first group of summary subjects and $P_{2_1}, P_{2_2}, ..., P_{2_n}$ is the second group. The degree of truth of summary (28) is evaluated with formula (29):

$$\begin{aligned} T(Q\ P_{1_1}, P_{1_2}, ..., P_{1_n}\ \textit{in comparison to}\ P_{2_1}, P_{2_2}, ..., P_{2_m}\ \textit{are}\ S_1) = \\ = \mu_Q \left(\frac{\sum_{i=1}^{n} \Sigma\textit{-count}(S_{1_{P_{1_i}}})}{\sum_{i=1}^{n} \Sigma\textit{ -count}(S_{1_{P_{1_i}}}) + \sum_{j=1}^{m} \Sigma\textit{ -count}(S_{1_{P_{2_j}}})} \right) \end{aligned} \tag{29}$$

The equation for degree of truth of type-2 summary:

$$\begin{aligned} T(\widetilde{Q}\ P_{1_1}, P_{1_2}, ..., P_{1_n}\ \textit{in comparison to}\ P_{2_1}, P_{2_2}, ..., P_{2_m}\ \textit{are}\ \widetilde{S}_1) = \\ = \mu_{\widetilde{Q}} \left(\frac{\sum_{i=1}^{n} \text{ card } (\widetilde{S}_{1_{P_i}})}{\sum_{i=1}^{n} \text{ card } (\widetilde{S}_{1_{P_i}}) + \sum_{j=1}^{m} \text{ card } (\widetilde{S}_{1_{P_j}})} \right) \end{aligned} \tag{30}$$

An example of such a summary is given:

$$\begin{aligned} &\textit{Larger part of teenage boys and school-age boys} \\ &\textit{than teenage girls and school-age girls is tall}\ [0.756] \end{aligned} \tag{31}$$

where $P_{1_1} =$ *teenage boys*, $P_{1_2} =$ *school-age boys*, $P_{2_1} =$ *teenage girls*, $P_{2_2} =$ *school-age girls* and $S_1 =$ *tall*.

Summaries in form (29) enable to work on two sets of subjects. The first set represents subjects that are being compared to subjects from second set. In the previous example, there are two subjects in each subject set which are distinguished from general subjects of boys and girls. This allows us to have a deeper look into large sets of subjects and group them into smaller and more precise summarization subsets.

4. Describing and summarizing data linguistically via multi-subject summaries: two application examples

Two application examples are presented. Both examples discover some dependencies between chosen groups of subjects using newly presented form of multi-subject linguistic summaries. There are two different databases used, one for each of two presented examples. For each database, a set of linguistic summaries is generated, sorted and selected with respect to the highest degree of truth. Finally, the results are related to those obtained via non-multi-subject forms, cf. (1) and (2) summaries, to compare original summarization methods to those introduced here.

The software created for testing purposes is based on the Java 1.7 SE platform.

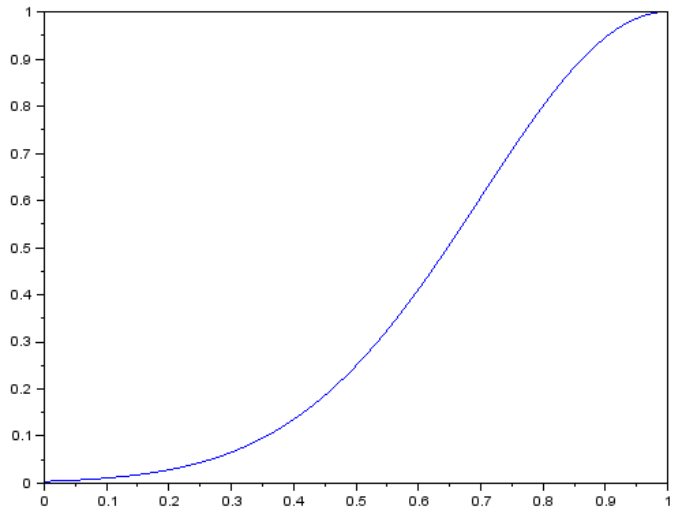

Figure 1. The membership function of the *MOST OF* linguistic quantifier.

4.1. Example 1: How age and gender of young people are weight related

The database used in this example contains data on children in the age of 7 up to 18 years old. The data describes e.g. children height, mass, date of birth, living conditions such as number of rooms in flat, number of people in family, family financial situation, etc. The database contain data on 13 956 children, including 6 991 boys and 6 965 girls.

The dataset is a real database collected by Department of Biostatistics, Medical University of Silesia. The experts of Department of Biostatistics have also assessed the results obtained by the proposed linguistic summarization methods as rewarding and providing reliable information although this information is semantically different than that provided by statistical analysis and methods. In their opinion, data analysis using linguistic summarizaion cannot replace statistics, but can help to describe large datasets using the most common way of people's communication: the natural language. From the point of view of soft computing and computational intelligence methods, the natural language descriptions generated by intelligent procedures are similar to human intuition, hence we can see these results as promising.

In the experiment, generated summaries are assumed to discover how children's age and gender is related to their height. Two subjects taken into account in multi-subject summaries are boys and girls. The process of logical splitting the database into two separated sets of data describing boys and girls, respectively, is exemplified by Table 1, 2 and 3, on Page 18). The relative quantifiers are used in the experiment called *most of, about two-third* and *about half* to represent the quantities in agreement for selected subjects, and to evaluate degrees of truth of the multi-subject summaries. The proposed membership functions for the quantifiers *most of* and *about two-third* are presented in Figure 1 and 2.

The generated summaries are based on qualifiers and summarizers represented by fuzzy sets. Sample summarizers and qualifiers are:

— tall (height)
— short (height)
— in early school age (age)
— teenager (age)

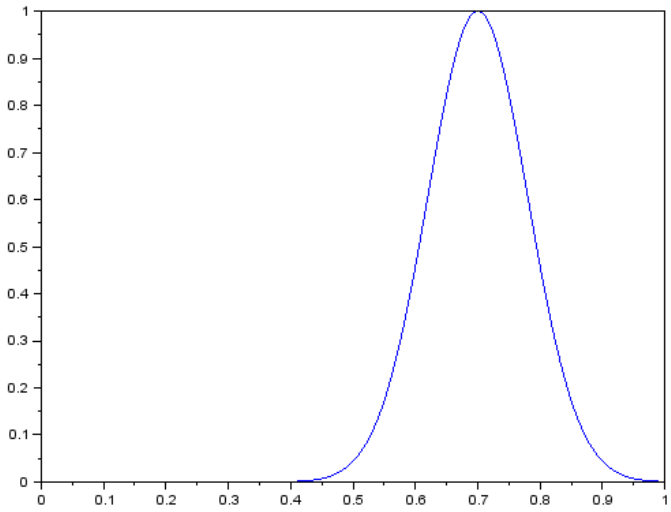

Figure 2. The membership function of the *ABOUT TWO-THIRD* linguistic quantifier.

The label *tall* is represented by fuzzy set $TALL$

$$TALL = \{\langle x, \mu_{TALL}(x)\rangle \colon x \in [150, 195], \mu_{TALL}(x) \in [0, 1]\} \tag{32}$$

where

$$\mu_{TALL}(x) = \begin{cases} \frac{(x-150)}{22.5}, & \text{if } 150 \leq x \leq 172.5 \\ \frac{(195-x)}{22.5}, & \text{if } 172.5 \leq x \leq 195 \\ 0, & \text{if } x \leq 150 \text{ or } x \geq 195 \end{cases} \tag{33}$$

the label *short* is represented by fuzzy set

$$SHORT = \{\langle x, \mu_{SHORT}(x)\rangle \colon x \in [103, 150], \mu_{SHORT}(x) \in [0, 1]\} \tag{34}$$

where

$$\mu_{SHORT}(x) = \begin{cases} \frac{(x-103)}{23.5}, & \text{if } 103 \leq x \leq 126.5 \\ \frac{(150-x)}{23.5}, & \text{if } 126.5 \leq x \leq 150 \\ 0, & \text{if } x \leq 103 \text{ or } x \geq 150 \end{cases} \tag{35}$$

Analogously, the label *teenage* is represented by fuzzy set

$$TEENAGE = \{\langle x, \mu_{TEENAGE}(x)\rangle \colon x \in [13, 18], \mu_{TEENAGE}(x) \in [0, 1]\} \tag{36}$$

where

$$\mu_{TEENAGE}(x) = \begin{cases} \frac{(x-13)}{2.5}, & \text{if } 13 \leq x \leq 15.5 \\ \frac{(18-x)}{2.5}, & \text{if } 15.5 \leq x \leq 18 \\ 0, & \text{if } x \leq 13 \text{ or } x \geq 18 \end{cases} \tag{37}$$

and the label *early school age* is represented by fuzzy set

$$\begin{aligned} &EARLY\ SCHOOL\ AGE = \\ &= \{\langle x, \mu_{\textit{EARLY SCHOOL AGE}}(x)\rangle \colon x \in [7, 12], \mu_{\textit{EARLY SCHOOL AGE}}(x) \in [0, 1]\} \end{aligned} \tag{38}$$

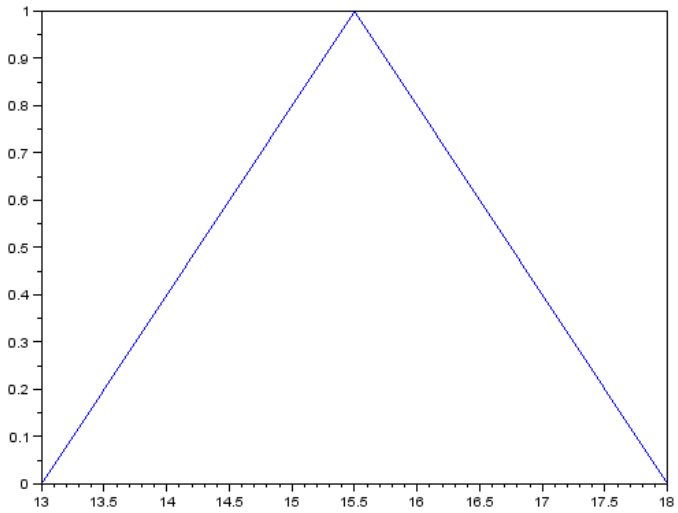

Figure 3. The membership function of the *TEENAGE* fuzzy set.

where

$$\mu_{\textit{EARLY SCHOOL AGE}}(x) = \begin{cases} \frac{(x-7)}{2.5}, & \text{if } 7 \leq x \leq 9.5 \\ \frac{(12-x)}{2.5}, & \text{if } 9.5 \leq x \leq 12 \\ 0, & \text{if } x \leq 7 \text{ or } x \geq 12 \end{cases} \tag{39}$$

The plot of the membership functions of *TEENAGE* is presented in Figure 3.

Results and interpretation The output of the experimental software, i.e. the generated summaries, are collected in Table 5. For each summary, the evaluated degree of truth (column T) and the form of the summary (column "Summary form"), are provided. The "Summary form" refers to the number of equation in this paper, that means (4), (14), (20), (24) refer to the first, the second, the third and the fourth form of a multi-subject linguistic summary given in Section 3, respectively, and (1) and (2) refer to the older forms of linguistic summaries.

According to expert' opinion, the results are intuitively correct. The first eight summaries 1.-8. are constructed according to the first form of a multi-subject linguistic summary (4). Analyzing their degrees of truth we can see that there is no disproportion between information on boys or girls. Looking at the summaries 1., 2. and 3. we can assume that the number of teenage boys should be similar to the number of teenage girls. Relatively, number of boys in early school age is more or less equal to the number of girls in the same age. Analysing seummaries 5.-7 the same can be assumed about subsets representing children's height. The only two values of T which are different from the rest in the first group represent summaries 4. and 8. This is related to the different quantifiers used in 4. and 8. Quantifier „about half" used in summary 4. shows clearly that if we divide data into early school age boys, early school age girls, teenage boys and teenage girls, all these subsest will be more or less equal.

The next summaries, 9.-16., lead us to the conclusion that there are more tall girls than tall boys in early school age: e.g. summary 9. contains the opposite statement, i.e. boys relatively to girls in the early school age are tall, and it is of the very low degree of truth. The situation changes for teenagers: there are more tall teenager boys than teenager girls, summary 10. Also, it cannot be said that in comparison to boys, major part of teenager girls are short, because it would mean that there are many teenager girls from 103cm to 150cm

Table 5. Sample multi-subject summaries illustrating relations between children age and height

No.	Summary	$[T]$	Summary form
1.	Most of girls in comparison to boys are in early school age	0.495	(4)
2.	Most of boys in comparison to girls are in early school age	0.505	
3.	Most of girls in comparison to boys are teenagers	0.511	
4.	About half of boys in comparison to girls are teenagers	0.994	
5.	Most of girls in comparison to boys are tall	0.206	
6.	Most of boys in comparison to girls are tall	0.298	
7.	Most of girls in comparison to boys are short	0.249	
8.	About two-thirds of boys in comparison to girls are short	0.043	
9.	Most of boys in comparison to girls being in early school age, are tall	0.004	(14)
10.	Most of boys in comparison to girls being teenagers, are tall	0.129	
11.	Most of girls in comparison to boys being in early school age, are short	0.124	
12.	About half of girls in comparison to boys being teenagers, are short	0	
13.	Most of girls being in early school age, in comparison to boys are short	0.101	(20)
14.	Most of girls being teenagers, in comparison to boys are short	0.004	
15.	Most of boys being teenagers, in comparison to girls are tall	0.098	
16.	About two-thirds of boys in early school age in comparison to girls, are tall	0	
17.	More boys than girls are tall	0.534	(24)
18.	More girls than boys are short	0.5	
19.	More boys than girls are teenagers	0.49	
20.	More girls than boys are teenagers	0.510	
21.	More boys than girls are in early school age	0.506	
22.	About half of children are girls	1	(1)
23.	Most of children are in early school age	0.32	
24.	About two-thirds of boys are tall	0	
25.	Most of boys being tall are teenagers	0.031	(2)

height summary 10. (the reader must take into consideration that children in the dataset was from 103cm to 195cm tall, so in this circumstances, a short child is more or less between 103cm and 150cm tall).

Summaries from 17. to 20. confirm lack of substantial disproportion between number of tall boys and tall girls and teenager boys and teenager girls. There are not many more tall boys than tall girls, according to summaries 17. and 18 and there are only a few more teenager girls than teenagres boys, according to summaries 19. and 20.. Summary 21. confirms that there are more teenager girls (the number of early school aged boys is slightly bigger than early school aged girls).

Using the older forms of the linguistic summaries, i.e. (1) and (2), provides us with extensive information about the analysed dataset. Extending summarizations set from Table 5 with summaries in known forms, 22.-25., completes our knowledge on the summarized database. For example, information on proportions between boys and girls, amount of tall boys, tall girls, teenager boys, teenager girls, teenage boys which are tall, early school aged girl which are short, are provided. The dedicated algorithm can evaluate degrees of truth, analyse summaries, select the best (the most informative) summaries and present in a clear and intuitive form, e.g. *About half of children are girls. Most of boys in comparison to girls are all. About two-third of girls being in early school age, in comparison to boys are tall*. In [21] the new approach of how to discard not promising summaries is presented, which can be used to select the most informative of them. Automated analysis of summaries can be done using the concept presented in [22]. The last conclusion shows in particular, that newly proposed multi-subject summaries of databases do not exclude the older forms, but can be used together

with them, to extend and improve the process of extracting and representing knowledge from large datasets.

4.2. Example 2: gender differences across the world

The database used in the second experiment contains data on male and female from around the world. The data describes e.g. level of education, age, monthly household income, country, population, etc. The database contain more than $150\ 000$ records.

In the experiment, generated summaries are assumed to discover associations between gender, age, education and monthly income accordingly to world region. The Multi-Subject linguistic summaries are generated here for the following sets of subjects:

— all female, all male
— female from Poland, male from Poland
— people from low income region, people from high income region

Subjects distinguished for N-Subjects summarization:
— female from Poland
— female from Finland
— female from Afghanistan
— female from Kenya

The process of logical splitting the database into separated sets of subjects is described in Section 2.

The relative quantifiers are used in the experiment called *less, about half* and *a lot of* to evaluate degrees of truth of the summaries. The proposed membership functions for the quantifiers *less*, *about half* and *a lot of* are presented in Figure 4, 5 and 6.

The generated summaries are based on qualifiers and summarizers represented by type-1 or type-2 fuzzy sets. Sample summarizers and qualifiers are:

— low (monthly income)
— high (monthly income)

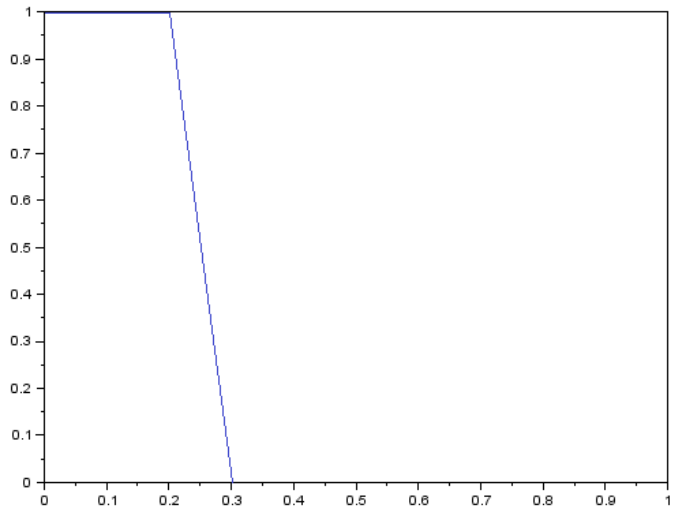

Figure 4. The membership function of the *LESS* linguistic quantifier.

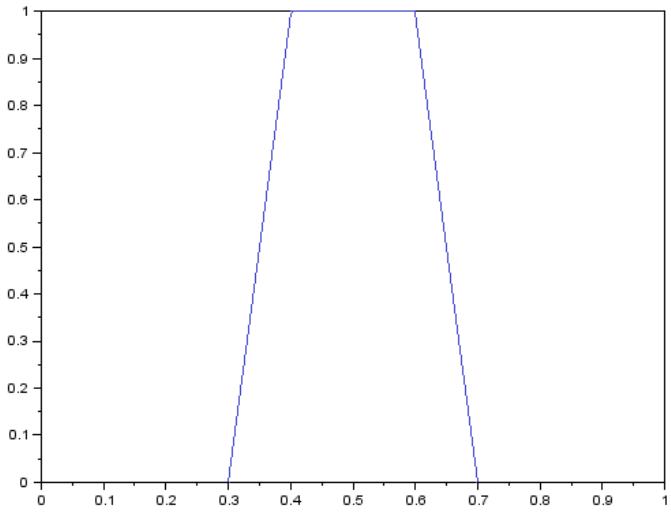

Figure 5. The membership function of the *ABOUT HALF* linguistic quantifier.

— productive (age)
— elderly (age)
— primary (education)
— tertiary (education)

Values for the monthly income attribute are represented by numbers from 0 to 5, are e.g. 0 means that there is no information about monthly income from a respondent, 1 means that respondent declared poorest household income, 3 means middle income and 5 is for richest income. For instance, the label *high income* is represented by fuzzy set $HIGH$

$$HIGH = \\ = \{\langle x, \mu_{HIGH}(x)\rangle \colon x \in [3,5], \mu_{HIGH}(x) \in [0,1]\} \tag{40}$$

where membership function have trapezoidal shape

$$\mu_{HIGH}(x) = \begin{cases} x-3, & \text{if } 3 \leq x \leq 4 \\ 1, & \text{if } 4 \leq x \leq 5 \\ 0, & \text{if } x \leq 3 \text{ or } x \geq 5 \end{cases} \tag{41}$$

The label *productive* age is represented by type-2 fuzzy set $\widetilde{P}$:

$$\widetilde{P} = \{\langle x, u_{productive}, \mu_x(u_{productive})\rangle \colon x \in [20,50], u_{productive} \in [0,1]\} \tag{42}$$

where

$$u_{productive} = \begin{cases} \frac{(x-20)}{15}, & \text{if } 20 \leq x \leq 35 \\ \frac{(50-x)}{15}, & \text{if } 35 \leq x \leq 50 \\ 0, & \text{if } x \leq 20 \text{ or } x \geq 50 \end{cases} \tag{43}$$

and

$$\mu_x(u_{productive}) = \exp\left(-\frac{1}{2}\left(\frac{u_{productive} - m(x)}{0.1}\right)^2\right) \tag{44}$$

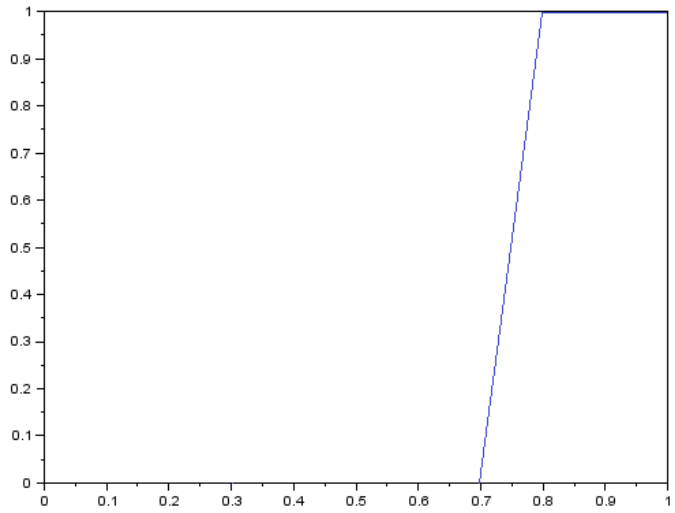

Figure 6. The membership function of the *A LOT OF* linguistic quantifier.

The label *ELDERLY* is represented by type-2 fuzzy set $\widetilde{E}$:

$$\widetilde{E} = \{\langle x, u_{elderly}, \mu_x(u_{elderly})\rangle : x \in [50, 80], u_{elderly} \in [0, 1]\} \tag{45}$$

where

$$u_{elderly} = \begin{cases} \frac{(x-50)}{15}, & \text{if } 50 \leq x \leq 65 \\ \frac{(80-x)}{15}, & \text{if } 65 \leq x \leq 80 \\ 0, & \text{if } x \leq 50 \text{ or } x \geq 80 \end{cases} \tag{46}$$

and

$$\mu_x(u_{elderly}) = \exp\left(-\frac{1}{2}\left(\frac{u_{elderly} - m(x)}{0.1}\right)^2\right) \tag{47}$$

Education level is a number from 0 to 3, where 0 means that there is no information about respondent education level, 1 means that respondent have primary education, 2 is for secondary education and 3 id for tertiary education level. Label *primary* is represented by fuzzy set $PRIMARY$

$$\begin{aligned} PRIMARY = \\ = \{\langle x, \mu_{PRIMARY}(x)\rangle : x \in [0, 2], \\ \mu_{PRIMARY}(x) \in [0, 1]\} \end{aligned} \tag{48}$$

where membership function have triangular shape

$$\mu_{PRIMARY}(x) = \begin{cases} (x), & \text{if } 0 \leq x \leq 1 \\ 2 - x, & \text{if } 1 \leq x \leq 2 \\ 0, & \text{if } x \leq 0 \text{ or } x \geq 2 \end{cases} \tag{49}$$

The label *tertiary* is represented by fuzzy set $TERTIARY$

$$\begin{aligned} TERTIARY = \\ = \{\langle x, \mu_{TERTIARY}(x)\rangle : x \in [2, 4], \\ \mu_{TERTIARY}(x) \in [0, 1]\} \end{aligned} \tag{50}$$

where membership function have triangular shape

$$\mu_{TERTIARY}(x) = \begin{cases} (x-2) & \text{if } 2 \leq x \leq 3 \\ 4-x, & \text{if } 3 \leq x \leq 4 \\ 0, & \text{if } x \leq 2 \text{ or } x \geq 4 \end{cases} \tag{51}$$

Results and interpretation The output of the experimental software, i.e. the generated summaries, are collected in Table 6. similarly to the previous example, for each summary, the evaluated degree of truth (column T) and the form of the summary (column "Summary form"). The "Summary form" refers to the number of equation in this paper.

The summaries presented in the table are chosen by their highest degrees of truth, or to emphasize some features or associations. Results are paired to show their logical accordance with human intuitive thinking. According to expert's opinion, the results are intuitively correct.

Summaries 1.-8. are constructed according to the first form of a multi-subject linguistic summary (4). First two summaries shows that if it is true that about half of all female, in comparison to all male, have lowest monthly income, it also must be true that about half of all male, in comparison to all female, have lowest monthly income. The same association is presented by summaries 3. and 4. Next two pairs of summaries (5.-6. and 7.-8.) are selected to check if opposite statements have similar degrees of truth, which is a desired situation. This means that if statement "about half of people from *low income region* in comparison to people from *high income region*, have *lowest monthly income*" is true, then the statement "about half of people from *high income region* in comparison to people from *low income region*, have *highest monthly income* should also be truth. Additionally, the following assumption can be stated on the base of summaries 1.-8.: respondents which live in low income regions declare poorest monthly household more often than respondents living in high income regions. The opposite financial situation is declared by respondents living in high income regions. Also, there are many more well educated respondents in high income regions.

Summaries 9.-14. are constructed on the same idea as previously discussed (with the same quantifier but two opposite statements), but using second form of multi-subject summaries (14). In summaries 13.-14., one can observe that using opposite quantifiers and opposite statements result in obtaining similar degrees of truth. According to summaries 13. and 14., respondents living in high income regions are more likely to have highest monthly household after graduation than respondents from low income regions.

The next two pairs of summaries, 15.-16. and 17.-18., use opposite quantifiers (like "a lot of" and "less") with the same statement, to show that degrees of truth have then opposite values. These summaries are build using the third form (20).

Summaries 19.-22. are build without quantifiers (form 24. Values of degrees of truth of paired summaries (summaries 19.-20. and 21.-22.) sum up to 1 which is intuitive. Summaries 19. and 20. suggest that there are more female than male respondents which declare tertiary education level. Summaries 21. and 22. confirm the previous conclusion that respondents from high income regions declares tertiary education level more often than respondents from low income regions.

N-subject summaries, 23.-24., are constructed with two sets of subjects, both sets contains two, separated subsets of subjects. In summary 23., first subject is represented by sub-subjects: female from Poland and female from Finland, second subject: female from Afghanistan and female from Kenya. This solution enables more sophisticated data explo-

Table 6. Sample multi-subject summaries illustrating gender differences across countries

No.	Summary	$[T]$	Summary form
1.	About half of female in comparison to male, have lowest monthly income	1.0	
2.	About half of male in comparison to female, have lowest monthly income	1.0	
3.	About half of female from Poland in comparison to male from Poland, have tertiary education	1.0	
4.	About half of male from Poland in comparison to female from Poland, have tertiary education	1.0	(4)
5.	About half of people from low income region in comparison to people from high income region, have lowest monthly income	1.0	
6.	About half of people from high income region in comparison to people from low income region, have highest monthly income	1.0	
7.	A lot of people from high income region in comparison to people from low income region, have tertiary education	1.0	
8.	A lot of people from low income region in comparison to people from high income region, have only primary education	0.949	
9.	A lot of female from Poland in comparison to male from Poland, being in productive age, have tertiary education	0	
10.	A lot of male from Poland in comparison to female from Poland, being in productive age, have tertiary education	0	(14)
11.	About half of male from Poland in comparison to female from Poland, being in elderly age, have tertiary education	1.0	
12.	About half of female from Poland in comparison to male from Poland, being in elderly age, have tertiary education	1.0	
13.	Less people from low income region in comparison to people from high income region, having tertiary education, have highest montlhy income	0.775	
14.	A lot of people from high income region in comparison to people from low income region, having tertiary education, have highest montlhy income	0.775	
15.	A lot of female from Poland, being in productive age, in comparison to male from Poland, have tertiary education	0	
16.	Less female from Poland, being in productive age, in comparison to male from Poland, have tertiary education	1.0	(20)
17.	A lot of people from low income region, having tertiary education, in comparison to people from high income region, have highest montlhy income	0	
18.	Less people from low income region, having tertiary education, in comparison to people from high income region, have highest montlhy income	1.0	
19.	More female from Poland in comparison to male from Poland, have tertiary education	0.671	(24)
20.	More male from Poland in comparison to female from Poland, have tertiary education	0.329	
21.	More people from high income region in comparison to people from low income region, have tertiary education	0.851	
22.	More people from low income region in comparison to people from high income region, have tertiary education	0.149	
23.	A lot of female from Poland and female from Finland, in comparison to female from Afghanistan and female from Kenya, have tertiary education	1.0	(28)
24.	A lot of female from Afghanistan and female from Kenya, in comparison to female from Poland and female from Finland, have tertiary education	0	

ration and deeper understanding of large datasets. Both summaries confirms that there are more female respondents with tertiary education among women from Poland and Finland, rather then respondents from Afghanistan and Kenya which is intuitively correct.

5. Conclusions

The goal of the research is to elaborate fuzzy-based methods that make it possible to describe contents of databases in a human-friendly manner as possible, preferably: with natural or quasi-natural language. In this paper, we present an original concept that extends the known methods of data linguistic summarization and representation: Multi-Subject Linguistic Summaries of relational databases. In particular, we put emphasis on new and more interesting forms of linguistic summaries, that discover associations between different groups of subjects within the same set of data, i.e. P_1, P_2, ..., and this is the *novum* of the paper. On the contrary, the older forms can handle one subject P only (for bibliographical references, see Section 1). The new forms of linguistic summaries are given by Equations (4), (14), (20), (24), and (28) in Section 3. We also provide the details of evaluating degrees of truth of the new summaries, in Section 3, too. From the point of view of an average user, the most important detail of the Multi-Subject Linguistic Summaries is that the output of the proposed method remains texts or messages composed by a human. Sample applications of Multi-Subject Linguistic Summaries to a system providing users with natural-language-information on a chosen set of data, are described in Section 4. Especially, Example 2 (Section 4.2) is worth noticing because higher order fuzzy logic is applied to represent linguistic terms appearing in summaries. We believe the proposals introduced here, i.e. describing more than one subject by a summary, may have potential to extend the summarization methods already known in the scientific literature.

References

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

[2] Zadeh, L. A.: A computational approach to fuzzy quantifiers in natural languages. Computers and Maths with Applications, 9, pp. 149–184, 1983.

[3] Yager, R. R., Ford, M., Canas, A. J.: An Approach To The Linguistic Summarization of Data. In: Proceedings of 3rd International Conference, Information Processing and Management of Uncertainty in Knowledge-Based System, Paris, France, pp. 456–468. 1990.

[4] George, R., Srikanth, R.: Data Summarization Using Genetic Algorithms and Fuzzy Logic. In: Herrera, F., Verdegay, J. L. (eds.), Genetic Algorithms and Soft Computing, pp. 599–611. Physica–Verlag, Heidelberg, 1996.

[5] Kacprzyk, J., Yager, R. R.: Linguistic summaries of data using fuzzy logic. International Journal of General Systems, 30, pp. 133–154, 2001.

[6] Kacprzyk, J., Yager, R. R., Zadrożny, S.: A fuzzy logic based approach to linguistic summaries of databases. International Journal of Applied Mathematics and Computer Sciences, 10, pp. 813–834, 2000.

[7] Yager, R. R.: On ordered weighted averaging operators in multicriteria decision making. IEEE Transactions on Systems, Man, and Cybernetics, 18, pp. 183–190, 1988.

[8] Niewiadomski, A.: Six new informativeness indices of data linguistic summaries. In: Szczepaniak, P. S., Wgrzyn Wolska, K. (eds.), Advances in Intelligent Web Mastering, pp. 254–259. Springer-Verlag, 2007.

[9] Niewiadomski, A.: News Generating via Fuzzy Summarization of Databases. Lecture Notes in Computer Science, 3831, pp. 419–429, 2006.

[10] Zadrożny, S.: Imprecise queries and linguistic summaries of databases. Academic Publishing House EXIT, Warsaw, 2006. (in Polish).
[11] Bosc, P., Pivert, O.: Fuzzy querying in conventional databases. In: Zadeh, L. A., Kacprzyk, J. (eds.), Fuzzy Logic for the Management of Uncertainty, pp. 645–671. Wiley, New York, 1992.
[12] Raschia, G., Mouaddib, N.: SAINTETIQ: a fuzzy set-based approach to database summarization. Fuzzy Sets and Systems, 129, pp. 137–162, 2002.
[13] Rasmussen, D., Yager, R. R.: A fuzzy SQL summary language for data discovery. In: Dubois, D., Prade, H., Yager, R. R. (eds.), Fuzzy Information Engineering: A Guided Tour of Application's, pp. 253–264. Wiley, New York, 1997.
[14] Srikanth, R., Agrawal, R.: Mining quantitative association rules in large relational databases. In: The 1996 ACM SIGMOD International Conference on Management of Data, pp. 1–12. 1996.
[15] Zadrozny, S., Nowacka, K.: Fuzzy information retrieval model revisited. Fuzzy Sets and Systems, 160(15), pp. 2173–2191, 2009.
[16] Codd, E. F.: A Relational Model of Data for Large Shared Data Banks. Communications of the ACM, 13(6), pp. 377–387, 1970.
[17] Niewiadomski, A.: A Type-2 Fuzzy Approach to Linguistic Summarization of Data. IEEE Transactions on Fuzzy Systems, 16(1), pp. 198–212, 2008.
[18] Niewiadomski, A.: On Finity, Countability, Cardinalities, And Cylindric Extensions of Type-2 Fuzzy Sets in Linguistic Summarization of Databases. IEEE Transactions on Fuzzy Systems, 18(3), pp. 532–545, 2010.
[19] Niewiadomski, A.: Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions. Academic Publishing House EXIT, 2008.
[20] Wu, D., Mendel, J. M.: Linguistic summarization using IFTHEN rules and interval type-2 fuzzy sets. IEEE Transactions on Fuzzy Systems, 19(1), pp. 136–151, 2011.
[21] Pilarski, D.: Linguistic summarization of databases with quantirius: a reduction algorithm for generated summaries. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 18, p. 305, 2010.
[22] A. Wilbik, J. M. K.: A distance metric for a space of linguistic summaries. Fuzzy Sets and Systems, 208, pp. 79–94, 2012.