DOI 10.34658/jacs.2019.27.1

JOURNAL OF APPLIED
COMPUTER SCIENCE
Vol. 27 No. 1 (2019), pp. 7-26

# Linguistic Summaries of Graph Databases in Customer Relationship Management (CRM)

**Monika Bartczak, Adam Niewiadomski**

*Faculty of Technical Physics, Information Technology and Applied Mathematics*
*Lodz University of Technology*
*ul. Wólczańska 215, 90-924 Łódź, Poland*
*214881@edu.p.lodz.pl, adam.niewiadomski@p.lodz.pl*

**Abstract.** *The paper concentrates on data models that differ from the traditional relational one by Codd (1970). In particular, we are interested in processing graph databases (graph datasets) without any pre-configured structure, in which graph nodes may represent different objects and graph edges – relations between them. In this approach, the linguistic summarization methods for graph datasets are introduced, and differences for these methods with respect to traditional relational approach are shown, commented and improved in comparison to the preceding proposition (Strobin, Niewiadomski, 2016). The novelty of the paper is mostly the new form for summaries: Multi-Subject linguistic summaries of graph databases, previously introduced for relational databases (Superson, 2018).*
**Keywords:** *graph data model, graph databases, linguistic summarization of graph datasets, customer relationship management, Neo4j, data mining, data, fuzzy representations of data.*

## 1. Introduction

The paper focuses on the problem of adapting computational methods within Artificial Intelligence (AI) to rapidly increasing amounts of data stored and processed in various repositories. Currently known and developed methods have to deal with analyzing and processing data, the number of which is expressed even in petabytes (PB, $10^{15}$B). Using traditional relational representation [1] and related software to process such amount of data is naturally limited mostly because of inefficiency of known methods and limited structure of the model. That is why collections of data are more and more frequently represented using other data models, including the graph model. In addition, to obtain comprehensive answers to users queries in a short (meaning: efficient) response time, it is important to represent results not only with real numbers and/or statistics, but, primarily, with natural or quasi-natural language, the most comprehensible manner of communication to a typical user. So, we deal with expressions like *About half of objects have very high value of attribute X*, e.g. [2, 3, 4]. They are renewed and adapted to process and represent knowledge in graph datasets. It causes that the problem of selecting the subject of the summary is not trivial and requires some new searching methods, while in relational databases a simple equality "Subject = set of records" takes place.

Hence, apart from selection methods for subjects of summaries in graph datasets, the paper introduces new forms of linguistic summaries: Multi-Subject Type Linguistic Summaries and some quality measures that enable assessments for generated summaries more precisely than those dedicated for a relational data model, formerly pointed by Bartczak in [5], in Section 5. Finally, illustrative examples of summarizing real graph datasets are shown and commented.

## 2. Motivation and Related Work

The inspiration to start the research in this topic has been given by the PhD Thesis by Strobin. However, the current research refers to it and enhances the former approach.

A relational model of data organization was originally published by Codd in 1970 [1]. Since then, the relation model of data organizing and representing is the most popular and frequently applied as the main standard in applications. Unfortunately, relational databases are not being adapted to still growing and evolving requirements of web applications, especially these operating on Big Data, with amounts of data expressed in petabytes, with dynamic temporal variability, partially structured or unstructured organization, and used by large numbers of users [6]. Famous examples of such applications include e.g. social media Facebook, LinkedIn, or database management systems like Apache. The so-called NoSQL databases, meaning *Not Only SQL*, exist since late sixties, but in fact never explored intensively as a research topic [7]. Nowadays, large and leading IT companies, e.g. Google and Amazon have contributed to increase the popularity of the idea of "Not Only SQL", mostly to find solutions to problems related to scalability and parallel processing of large amounts of data, like those used in applications Gmail, Google Maps, Google Earth, Google Finance, and Google applications and search engine built [8].

In particular, in this paper, we focus on using artificial intelligence and fuzzy computing methods to mine and represent data collected in such sets organized within graph structures. We concentrate mostly on the linguistic representation of data stored in graphs with linguistic summaries of data, see [9]. Some former work have been done by the authors, e.g. representing and evaluating quality of data summaries via general and interval-valued fuzzy sets [3, 10] and [11], respectively. Besides, first attempts of summarizing partially structured data in graphs are presented in [6, 12]. Multisubjectivity in linguistic summaries is introduced in [13]. One of objectives of this paper is to test empirically the performance of graph database by mining and summarizing huge amounts of data. The problem of choosing subjects of summary in graph datasets is also addressed. Verification of calculations on the base of membership functions of fuzzy sets representing summarizers $S$ and other elements of summaries, like $Q$ – linguistic quantifiers, are presented. As the main novelty of the paper the authors propose a new method of representing data in graph model: Multi-Subject Type Linguistic Summaries of Graph Databases, formerly mentioned by Bartczak in [5].

Computational examples of graph datasets linguistic summarization are presented with a real database containing approx. 330 000 records using Neo4j DBMS.

# 3. Data structure and representation in graph databases

This section provides a brief description of the data structures in the representation of graphs and presenting possible ways to construct a graph database. In addition, it shows the differences and similarities of the relational model with graph model.

**Basics** Graph database use graph structure with nodes. In this model, the data objects are represented by vertices, and the links between them - by the edges.

Vertices can have any attributes, and can be freely associated [6]. Implementation of such a graph model is i.e. hierarchical structure in the company. The solution may be based on a database such as Neo4j, FlockDB [14].

The following is an example of which is shown illustration model graph database. The crimson color is indicated edges and the gray tops of the graph. This example is a model graph labeled. This means that the edges and vertices are labeled[1].

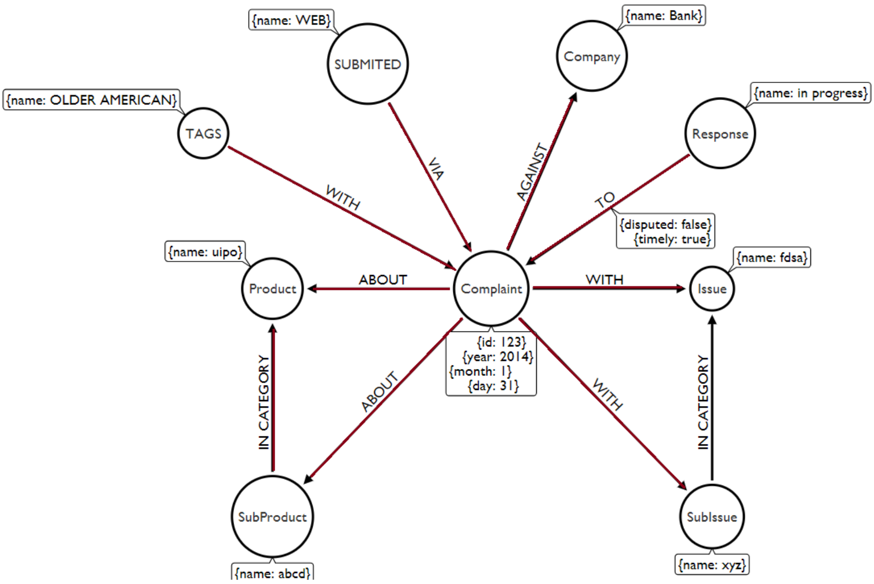


Figure 1: An example of a non-relational database: labeled graph model

The dataset organized within this model is used in the experiments, as shown in Figure 2.

Schema has been created based on the data contained in the database used. The data were collected in a format .csv file. Neo4j also has the ability to import data created earlier structure JSON format. I believe that with such a structure created I get interesting, rewarding me the results.

Schema generation is also possible by converting the relational model to a graph. You need to prepare a relational database schema, and then use 2G, which

[1]The structure of the graph database has been developed with the online program available at `http://www.apcjones.com/arrows/`.

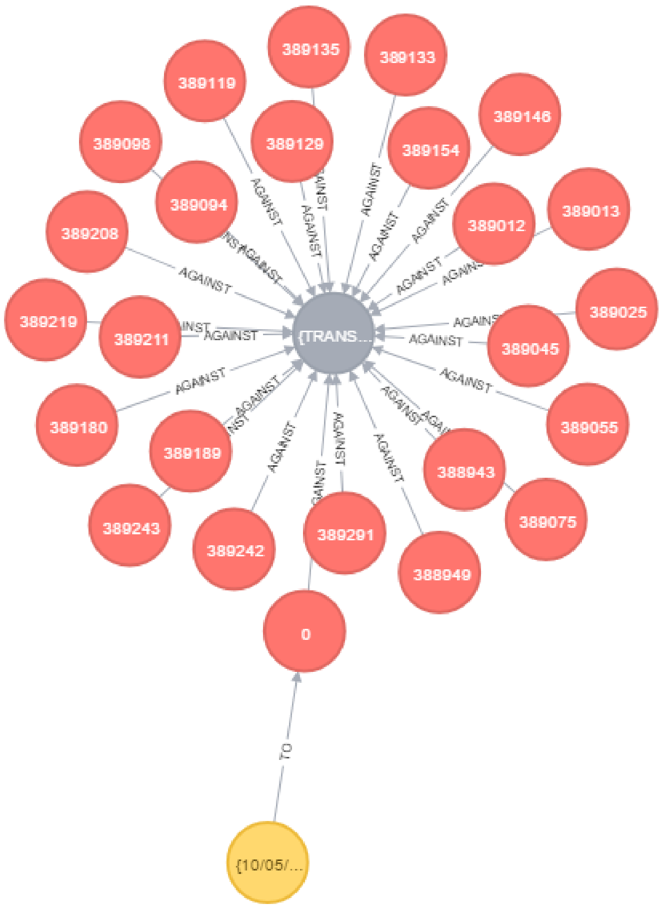


Figure 2: The graph model database used in the experiment (fragment). The database is generated with Neo4j.

performs the conversion of the relational model to graph model called "3NF Equivalent Graph". The transformation between the base and the relational graph is asymmetrical, i.e. the relational model to graph model is possible, but with the relational model to graph model.

**Notes on the drawbacks of a relational data model for data mining and representation** In this part of the article will be compared with the relational databases graph by combining their advantages and disadvantages.

A relational database has the following advantages:

- The ability to model any data structure without introducing redundancy or loss of precision, through a process of standardization;
- Ability to add, modify, and retrieve data using SQL (Structured Query Language).

Unfortunately, relational data model, however, has several disadvantages:

- Poor performance of complex queries;
- Low scalability difficulty in changing the structure of the database after a certain time;
- Relational model requires the creation of so-called ORM layers (Object - Relational Mapping) during cooperation with the technology and object-oriented languages.

Graph database eliminates problems relational model. Is the basis of efficient, flexible and easily scalable [15]. Performance model in search of the complex relationships between objects can argue the direct links between objects - the edges of the graph. Experiments have shown in [16] the article the graph database advantage over a relational database for queries complex (i.e. those which require the implementation of many operations splice) [6].

# 4. The linguistic summaries of graph databases

In this part of the publication will be demonstrated in the basic form of linguistic summary, as well as will be discussed and imaged several examples of calculating membership summarizer. Moreover, the pattern will be discussed in the Summary of Dependencies by Strobin.

## 4.1. The concept of linguistic summaries

The concept of linguistic summaries of databases is based on the Zadeh calculation. These calculation concerns expressions quantified linguistically. Two forms of linguistic summaries will be presented:

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

Ex. A lot of complaints from clients concerns a high amount [0.83]

$$Q \ \ P \ \ which \ \ W \ \ is/are \ \ S[T] \tag{2}$$

Ex. A lot of complaints from clients which made with the summer concerns a high amount [0.63].

These forms were presented among others in [17], [4] articles.

In both forms of $Q$ is a linguistic quantifier represented by an aggregation operator (in the present case is the fuzzy quantifier), for example, many, less, more than 900. $P$ is the subject of the summary, for example, men, women, car or cars. $S$ is summarizers, linguistic expressions, which refers to properties of objects involved in the summary. Examples summarizers in fuzzy sets high (increase) or low (increase) etc. $T$ $[0, 1]$ is a degree of truth and it determines how a defined formulation is close to reality. $T$ values are determined on the basis to the Zadeh calculus concerns quantifiable expressions linguistically and other methods described in the works. Symbol $W$ appears only in the second formula, it is a qualifier, represented by a fuzzy set that represents the additional properties of the objects involved in the summary [13].

### 4.2. The choice of an entity of the summary in the graph model

In the case of a relational database, the subject is a column. Linguistic summaries of the relational model are trivial. Unfortunately, in the case of graph model choice of entity is not obvious. In the graph $G$ with a set of vertices, in $V$ hould define the type of vertex to the linguistic summary, it means defined a subset of vertices $P$, which will be treated as subjects summaries. This is a defined condition $K$, which must take into consideration vertex, so he could find yourself in a set of subjects.

$$P = \{p \in V : K(p)\} \tag{3}$$

where $V$- is set of vertices in the graph $G$, $P$ is a set of subjects, $K$ is a condition which must take into consideration vertex in graph.

I choose in the analyzed case as an entity is complaints (problems) submitted by customers.

### 4.3. Summary of Dependencies: Original and Improved

The method of linguistic summaries of graphical relations introduced by Łukasz Strobin in [18] is the so-called Summary of Dependencies. The purpose of the new form is to answer the question: "What characterizes pairs of connected vertices?". **The subject** $P$ of this new form of summaries is not a set of objects and their attributes, but a set of pairs of vertices connected with each other by a given semantic dependence. We are changing the field of summaries. In the case of summaries of relationships between vertices, we can determine the **full set of** $P^G$ entities as a set of all vertex pairs **in the graph** $G$ between which the given semantic dependence occurs.

For such summaries, the **summarizers** $S$ concerns the relationship between vertex attributes. An example of such a summary of the database of people may be, for example, **"Most (**$Q$**) people (**$P^G$**) in married couples (**$Z^G$**) have a similar age (**$S$**)"**. As you can see, the quantifier Q has the same meaning as in the case of typical linguistic summaries. **The summarizers** $S$ can take the form of a fuzzy set (as in the example shown) or a sharp one, e.g. "the same age". $Z^G$ is the type of connection that must be present in each pair from the subject, that is - the set semantic dependence.

After selecting the set of vertex pairs $p_i = (V_i, V_k) \in P^G, i \neq j$ being the summation entity, attributes should be retrieved for each vertex. In general, the set of attributes for each vertex may be different, i.e. $A^{V_j} \neq A^{V_k}$. However, in the case of a summarizer operating on pairs of values, no value for attribute does not make sense, so the realm on which such summarizer operate is defined as $A = A_{V_j} \cap A_{V_k}$. In this proposed form of linguistic summaries, summarizers are functions that operate on two arguments - attribute values for each of the vertices of the pair constituting the summary entity. The first step is to define a function that compares the attribute values, and then - **to calculate the value of the membership function** $\mu$ **for a given linguistic variable** [18].

The following are two comparison functions, for discrete values of the attribute (where the value can be a set), and for numerical values:

1. The comparison function for the discrete attribute values:

$$sim^z(a(V_j), a(V_k)) = \frac{2 * |a(V_j) \cap a(V_k)|}{|a(V_j)| + |a(V_k)|} \tag{4}$$

Note that when a $(V_j)$ and a $(V_k)$ are collections function value is 1 when the values are the same, and 0 - if they are different. Thus, the above formula is a generalization of a simple equivalence.

**Example 1** *Let A - responses from the company to the consumer*
$(V_j)$ *= {"Closed with monetary relief", "Closed", "Closed with explanation"}*
$(V_k)$ *= {"Closed without relief", "Closed with monetary relief"}*

$$sim^z = \frac{2*|\{\text{"closed with.","closed","with explanation"}\} \cap \{\text{"closed without.","closed with."}\}|}{|\{\text{"closed with.","closed","with explanation"}\}|+|\{\text{"closed without.","closed with."}\}|} =$$
$$= \frac{2*1}{3+2} = 0.4$$

2. The comparison function for the numerical values:

$$sim^z(a(V_j), a(V_k)) = \frac{2 * |a(V_j) - a(V_k)|}{|a(V_j)| + |a(V_k)|} \tag{5}$$

**Example 2** *Let a - the number of complaints,* $(V_j) = 15000$ *and* $(V_k) = 2000)$. *Then*

$$sim^z = 1 - \frac{2*|15000-2000|}{|15000|+|2000|} = 1 - \frac{2*13000}{17000} = 1 - \frac{26000}{17000} \approx 1 - 1.52 \approx -0.52$$

After analysis, I can say that the model is modified. When using the above formula in the calculation comparison function attribute values, in some cases we receive negative values. What we can see in the example above. After several attempts to modify and adapt the model obtained the following form:

$$sim^z(a(V_j), a(V_k)) = 1 - \frac{|a(V_j) - a(V_k)|}{|a(V_j)| + |a(V_k)|} \tag{6}$$

Then, for the above example, we obtain the following result:

$$sim^z = 1 - \frac{|15000-2000|}{|15000|+|2000|} = 1 - \frac{13000}{17000} = \frac{4000}{17000} \approx 0.235$$

Both functions operate in $sim^z$ set $[0, 1]$, so that it is possible to apply one group of linguistic variables, i.e. "The same", "very like", "similar", "somewhat similar ", "different".

In addition to linguistic variables such it is also possible to define other, eg. on dates. Then the comparison function ($sim^z$) can be eg. The difference between dates in years, and properly selected must be linguistic variables.

Using the above formulas ((3) and (4)), express quality measures of linguistic summaries just like classic summaries. The difference lies in the field of the summary and is not a set of tuples, but the set of pairs of vertices.

For such defined summaries, we define the following quality measures:

**The degree of truth $T$ for a summary of dependencies is evaluated with a formula**

$$T(QP^G \ \ are \ \ S_j^z) = \mu_Q \frac{r^G}{m} \tag{7}$$

where $m$ **- number of pairs of vertices** which are summaries subjects and $r^G$ **is expressed by the** formula:

$$r^G = \sum_{i=1}^{m} \mu_S(sim^z(v_j, v_k)) \tag{8}$$

Where $sim^z$ is a function of comparing the attribute values, and $\mu_S$ - membership function for the summarizer $S$. Similarly to the classic linguistic summaries, in summaries, we define the summary of dependencies with a qualifier [18].

$$r^G = \frac{\sum_{i=1}^{m}(\mu_{S^Z} \wedge W^z(P_i^G))}{\sum_{i=1}^{m} \mu_{W^z}(P_i^G)} = \frac{\sum_{i=1}^{m} \mu_S \wedge W^z(sim^z(v_j, v_k))}{\sum_{i=1}^{m} \mu_{W^z}(sim^z(v_j, v_k))} \tag{9}$$

Linguistic summaries frequently refer to the whole set, for example, "A lot of complaints from clients concerns a high amount." A qualifier causes that we refer to the appropriate subject, for example. "A lot of complaints from clients which made with the summer concerns a high amount.".

# 5. Multi-Subject Type Linguistic Summaries of Graph Databases

Considering the formula (6), the author of the research work has developed a new way to calculate the degree of truth of the linguistic summaries in the graph data model. According to the author of the thesis, the new formula is simpler to implement and more transparent.

It includes two forms: first, where there is no linguistic quantifier and the second one in which it does exist. Then on their basis invented and introduced an original and new method - multi-subject type linguistic summaries of graph databases.

The newly defined formula takes into account the $n$ number of the subject. This can be seen in the following examples.

The first of the proposed forms for calculating the degree of truth of the summary for one subject may take the following formula:

$$T = 1 - \frac{|V_j - V_k|}{|V_j + V_k|} \tag{10}$$

where $T$ is degree of truth, $V_j$ and $V_k$ are vertices in the graph $G$. $V_j$ is count of data, in which an entity meets selected condition (has the desired feature), and $V_k$ is a count of data, which does not meet this condition. This can be shown with mathematical properties: $V_j \in V_k$, $V_k \notin S$ and $V_j^d = V_k$. In the analyzed case $S$ is summarizers, which is represented through a fuzzy set or classical set.

When calculating the degree of truth of the summary, at the final step we subtract the obtained result based on checking the similarity of attributes from 1. We do this, that to get the result which showing us how close true is summary. Otherwise, we will get information on how the summary is far from the truth.

**Example 3** *Let $V_j$ will be the complaints which got a timely response, and $V_k$ are complaints which not got a timely response. t was assumed that the complaints that received responses in the time accepted were* 1500*, and complaints that did not get them* 200.

*Summary: Complaints have been dealt with at a specific time (timely response) [0.82].*

$$T = 1 - \frac{|1500-200|}{|1500+200|} = 1 - \frac{|1300|}{|1700|} = \frac{1400}{1700} \approx 0,82$$

Degree of the truth of the summary is 0,82 the sentence (summary) was considered to be true.

Degree of truth for second with mentioned above (with a linguistic quantifier) form may be evaluated with a formula:

$$T = 1 - \left| \frac{(R_{MAX} + R_{MIN})}{2} - \left(1 - \frac{|V_j - V_k|}{|V_j + V_k|}\right) \right|, \tag{11}$$

where $R_{MAX}$ is the maximum range and $R_{MIN}$ is the minimum range of the linguistic quantifier. In this form, expert knowledge is required. Exemplary a frequency distribution is presented below:

$x \in < 0; 0.1)$ *very little*

$x \in < 0.05; 0.45)$ *few*
$x \in < 0.40; 0.60)$ *about half*
$x \in < 0.55; 0.85)$ *many*
$x \in < 0.80; 1.00 >$ *most*
*where* $x \in X$
Quantifier $Q$ can be represented in this way: $Q \in < R_{MIN}, R_{MAX} >$.

**Example 4** *Let $V_J$ will be the complaints which got a timely response, and $V_K$ are complaints which not got a timely response. It was assumed that the complaints that received responses in the time accepted were* 1500*, and complaints that did not get them* 200*.*

*Summary: Many complaints have been dealt with at a specific time (timely response) [0.88].*

$$T = 1 - \left| \frac{(0.85+0.55)}{2} - \left(1 - \frac{|1500-200|}{|1500+200|}\right) \right| \approx 1 - 0.12 \approx 0.88$$

Degree of the truth of the summary is $0,88$ – the sentence (summary) was considered to be true.

A new method has been proposed for linguistic summaries in the graph data model: Multi-Subject Type Linguistic Summaries. Multi-subject linguistic summaries are an extension of the classic, existing concepts of summaries databases. Multi-subject type for linguistic summaries allows of summaries the data based on not just one entity. Until now, they did not exist for a graph data model.

The author's method takes four forms:

The first of the presented forms for multi-subject type linguistic summaries:

$$QP_1 \ \ relatively \ \ to \ \ P_2 \ \ is/are \ \ S_1 \tag{46}$$

where $S_1$ is summarizers which are representing through fuzzy set, and $P_1$ and $P_2$ are linguistics subjects. In graph databases, subjects are vertices, which are shown through symbols $V_j$ and $V_k$. Degree of truth in this form is calculating through a formula:

$$T = 1 - \frac{|(V_j + V_k) - (V_j + V_k)|}{|(V_j + V_k) + (V_j + V_k)|}| \tag{12}$$

**Example 5** *Ex. summary: Complaints made by older Americans relative to complaints made by elderly people receive a timely response.*

*Count of complaints made by older Americans is* 1500*, count of complaints no* *made by older Americans is* 500*, count of complaints made by elderly people is* 200*, and count of complaints no made by elderly people is* 1700*.*

$$T = 1 - \frac{|(1500+500)-(200+1700)|}{|(1500+500)+(200+1700)|} =$$
$$= \frac{38}{39} \approx 0,97$$

Degree of the truth of the summary is 0,97 - the sentence (summary) was considered to be true.

The second of the presented forms for multi-subject type linguistic summaries:

$$P_1 \;\; or \;\; P_2 \;\; is/are \;\; S_2 \tag{13}$$

Calculation degree of truth it possible by applying the following formula:

$$T = 1 - \frac{|V_j + V_k|}{\sum P^G} \tag{14}$$

where $P^G$ is count of full set of subject, set of all vertex pairs in the graph $G$. $V_j$ is count of data, in which the first subject meets the selected condition (subject has a feature), and $V_k$ is count of data, in which the second subject meets this condition.

**Example 6** *Ex. summary: Complaints made by older Americans or complaints made by elderly people receive a timely response.*

*Let as in the previous example: count of complaints made by older Americans is* 1500*, count of complaints made by elderly people is* 200*. The full set of entities is* 5000*.*
*Then:*

$$T = 1 - \frac{|1500+200|}{|5000|}$$
$$T = 1 - \frac{|1700|}{|5000|}$$
$$T = 1 - \frac{|17|}{|50|} \approx 0,66$$

Degree of the truth of the summary is 0,66 - the sentence (summary) was considered to be true.

The first and second forms do not require the use of measures or fuzzy models e.g. quantifiers to compare two subjects. This approach allows you to quickly generate summaries whose content is intuitive.

For the third of the proposed forms for the set of data which were represented by the graph database, calculation the degree of truth it is possible by applying the following formula:

$$Q \ P_1 \ relatively \ to \ P_2 \ is/are \ S_1 \tag{15}$$

$$T = 1 - \left| \frac{(R_{MAX} + R_{MIN})}{2} - \left(1 - \frac{|(V_j + V_k) - (V_j + V_k)|}{|(V_j + V_k) + (V_j + V_k)|}\right) \right| \tag{16}$$

**Example 7** *An example of a summary that uses the third form (15): Most complaints made by older Americans with reference to the elderly people are dealt with in a timely accepted [0.93].*

*Let: complaints made by older Americans constitute* 1500*, omplaints not filed by elderly Americans* 500*, complaints made by elderly people* 200*, and complaints not submitted by elderly people* 1700*. The linguistic quantifier "most" accepts values in the range* $x \in < 0.8; 1 >$.
*Then*

$$T = 1 - \left| \tfrac{(0.8+1)}{2} - \left(1 - \tfrac{|(1500+500)-(200+1700)|}{|(1500+500)+(200+1700)|}\right) \right|$$
$$T = 1 - \left| 0.9 - \left(1 - \tfrac{|100|}{|3900|}\right) \right| \approx 0.93$$

Degree of the truth of the summary is 0.93 – the sentence is considered to be true.

Summaries in the third form allow receiving information about selected features $S_1$ subjects, depending on the conditions that both subjects should be meets. In this case, will be taken on attention data which concerns older Americans and elderly people.

The fourth form among the suggested forms of multi-subject type linguistic summaries on graph database looks like this:

$$QP_1 \ or \ P_2 \ is/are \ S_1 \tag{17}$$

Degree of truth can be evaluated by the formula:

$$T = 1 - \left| \frac{(R_{MAX} + R_{MIN})}{2} - \left(1 - \frac{|V_j + V_k|}{\sum P^G}\right) \right| \tag{18}$$

**Example 8** *Ex. summary: Most of the complaints made by older Americans or elderly people received a timely response.*

*Let: complaints made by older Americans is* 1500*, and complaints made by*

*elderly people is* 200*. Quantifier „most"has values in a range* $< 0.8, 1 >$.
*Then:*

$$T = 1 - \left|\frac{(0{,}8+1)}{2} - \left(1 - \frac{|1500+200|}{5000}\right)\right|$$
$$T \approx 0,76$$

Degree of the truth of the summary is 0,76 – the sentence is considered to be true.

Summaries in the fourth form allow receiving information about selected features $S_1$ subjects, depending on the conditions that one of the subjects must be meets (at least one). In this case, will be taken on attention data which concerns older Americans or elderly people.

In this chapter, I discussed four selected forms of linguistic summaries. I presented a few forms to prove and illustrate that they are possible linguistic summaries for more than one subject of graph databases. Forms of linguistic summaries can be much more. I analyzed the forms proposed in [19] for graph database

$$QP_1 \ \ relatively \ \ to \ \ P_2 \ \ is/are \ \ S_1 \tag{19}$$

$$QP_1 \ \ being \ \ S_2 \ \ relatively \ \ to \ \ P_2 \ \ is/are \ \ S_1 \tag{20}$$

$$More \ \ P_1 \ \ than \ \ P_2 \ \ is/are \ \ S_1 \tag{21}$$

and I created a new form:

$$QP_1 \ \ or \ \ P_2 \ \ being \ \ S_2 \ \ is/are \ \ S_1 \tag{22}$$

The difference between multi-subject type linguistic summaries of a graph database and relational database is data selection.

# 6. Results and comments

The results of the degree of truth of linguistic summaries of a graph database, calculated by my application are presented below in the Table 1. For each summary, the evaluated degree of truth (column T) and the form of the summary (column "Summary form"), are provided.

Table 1: Sample summaries with evaluated a degree of truth for different summary forms.

| No. | Summary | T | Summary form |
|---|---|---|---|
| 1. | Complaints which have been made by older Americans received a timely response | **0,99** | Summary of Dependencies |
| 2. | Complaints which have been made by older Americans received a timely response for the benefit of the customer | **0,99** | Summary of Dependencies |
| 3. | Complaints which have been made by younger people received a response for without benefit of the customer | **0,99** | Summary of Dependencies |
| 4. | In summer, many of complaints which have been made by military received a timely response | 0,80 | Summary of Dependencies |
| 5. | A few complaints which have been made by the military received a response for benefit of the customer | 0,88 | Linguistic summaries for one subject of graph databases: author's method |
| 6. | Most of the complaints which have been made by elderly Americans have been sent by phone | 0,68 | Linguistic summaries for one subject of graph databases: author's method |
| 7. | Very few complaints have been made throughout the phone | 0,73 | Linguistic summaries for one subject of graph databases: author's method |
| 8. | In the summer, about half of the complaints were sent by the client over the network | 0,79 | Linguistic summaries for one subject of graph databases: author's method |
| 9. | In winter a few complaints made by the clients received a timely response with benefit for clients. | 0,25 | Linguistic summaries for one subject of graph databases: author's method |
| 10. | In the summer many complaints made by the customers received a timely response with benefit for customers. | 0,80 | Linguistic summaries for one subject of graph databases: author's method |
| 11. | In summer many complaints made by older Americans or elderly people received a timely response with the benefit for clients. | 0,19 | Author's method: multi-subject type linguistic summaries of graph databases |
| 12. | In the spring a few complaints made by younger people or grown-up society were sent by Web | 0,24 | Author's method: multi-subject type linguistic summaries of graph databases |
| 13. | In spring a few complaints made by younger people over the Web compared to grown-up society | 0,35 | Author's method: multi-subject type linguistic summaries of graph databases |

According to expert opinion, the results are intuitively correct.

Newly proposed multi-subject summaries of graph databases do not exclude the older forms but can be used together with them, to improve and extend the process of extracting and representing knowledge from large datasets.

For the sentences "Complaints which have been made by older Americans received a timely response", "Complaints which have been made by older Americans received a timely response for the benefit of the customer" and "Complaints which have been made by younger people received a response for without benefit of the customer" achieved the greatest degree of truth of 0,99.

```
timelyTrue:19388 timelyFalse:332
W okresie wakacyjnym większość zażaleń zostało rozpatrzonych  na korzyść klienta na czas 0,99151068835021
W okresie wakacyjnym większość zażaleń zostało rozpatrzonych  na korzyść klienta na czas True
```

Figure 3: Example summary "In summer, most of the complaints made by customers received a timely response with benefit for clients" with calculated the degree of truth [0,99]. The sentence is considered to be true (TRUE).

where

- Complaints with the holiday period **a qualifier** (*W*). For the holiday period, they were adopted months: June, July, and August,
- the most – **linguistic quantifier** (*Q*). Expert assumed that most it more than 0.80,
- complaint - **subject** (*P*),
- with benefit - **a quantifier** (*Q*). Selected were answers that were considered with benefit for the customer, for example. "Culminating with monetary relief" or "completed with the help of non-pecuniary."

## 7. Conclusions and future work

One purpose of the publication was to verify the effectiveness and efficiency of the graph model by performing operations on a large number of data. The average response time for any of the queries for the database including $3,31,892$ vertices, edges and $1,414,085$ and $329,304$ labels lasted about 2 ms. In the case of very complex queries, they obtained a quick response. It can be concluded that the graph model is very efficient.

Another of the goals was to check the current state of knowledge in this field. Unfortunately, it is small. The above-mentioned doctoral [18] dissertation is currently one of the few sources of knowledge on the subject.

The use of linguistic summaries significantly improves knowledge management. This is a result of the transfer of a clear, brief and understandable information for all.

It has been proven that the Summary of Dependencies has been implemented and applied in an appropriate way.Analyzed, it established and created a new method of summary graph databases: Multi-Subject Type Linguistic Summaries of graph databases. It has been found empirically that it is possible to use multi-subject type linguistic summaries in graph databases.

It is possible to further pursue studies through:

- use of other quality measures of linguistic summaries in graph databases other than yet;
- execution linguistic summaries on another non relational database.

# References


[1] Codd, E. F., *A Relational Model of Data for Large Shared Data Banks*, Communications of the ACM, Vol. 13, No. 6, 1970, pp. 377–387.

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

[3] Niewiadomski, A., *Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions*, Academic Publishing House EXIT, 2008.

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

[5] Bartczak, M., *Podsumowania lingwistyczne grafowych baz danych w zarządzaniu relacjami z klientem (CRM)*, Master's thesis, Instytut Informatyki, Politechnika Łódzka, 2018, (in Polish).

[6] Strobin, L. and Niewiadomski, A., *Linguistic Summaries of Graph Datasets Using Ontologies: An Application to Semantic Web*, In: Computational Collective Intelligence 7th International Conference ICCCI 2015, 2015, pp. 380–389.

[7] *NoSQL*, http://www.leavcom.com/pdf/NoSQL.pdf, Accessed: 07.12.2019.

[8] Chang, F., Dean, J., Ghemawat, S., Hsieh, W., Wallach, D., Burrows, M., Chandra, T., Fikes, A., and Gruber, R., *Bigtable:A Distributed Storage System for Structured Data*, https://static.googleusercontent.com/media/research.google.com/pl//archive/bigtable-osdi06.pdf, Accessed: 07.12.2019.

[9] Kacprzyk, J., Yager, R. R., and Zadrożny, S., *Fuzzy linguistic summaries of databases for an efficient business data analysis and decision support*, In: Knowledge Discovery for Business Information Systems, edited by W. Abramowicz and J. Żurada, Kluwer Academic Publisher, B. V. and Boston, 2001, pp. 129–152.

[10] Niewiadomski, A., *A type-2 fuzzy approach to linguistic summarization of data*, IEEE Transactions on Fuzzy Systems, Vol. 16, No. 1, 2008, pp. 198–213.

[11] Niewiadomski, A., Ochelska, J., and Szczepaniak, P. S., *Interval-valued linguistic summaries of databases*, Control and Cybernetics, Vol. 35, No. 2, 2006, pp. 415–444.

[12] Strobin, L. and Niewiadomski, A., *Integration of Multiple Graph Datasets and Their Linguistic Summaries: An Application to Linked Data*, In: Artificial Intelligence and Soft Computing 2016, Part I, 2016, pp. 333–343.

[13] Niewiadomski, A. and Superson, I., *Multi-Subject Type-2 Linguistic Summaries of Relational Databases*, In: Frontiers of higher order Fuzzy Sets, edited by A. Sadeghian and H. Tahayori, Springer-Verlag, 2015.

[14] *NoSQL*, https://www.grupa-tense.pl/blog/czy-nosql-to-przyszlosc-baz-danych, Accessed: 07.12.2019.

[15] Słotwiński, D., *Grafowe bazy danych – przegląd technologii*, https://ai.ia.agh.edu.pl/_media/pl:dydaktyka:ztb:2010:projekty:gdb:grafowe_bazy_danych.pdf, 2010, (in Polish).

[16] Strobin, L. and Niewiadomski, A., *Wielokrotne przyspieszenie dzialania aplikacji poprzez zastosowanie technologii nierelacyjnych baz danych*, Economic Studies. University of Economics in Katowice, , No. 199, 2014, (in Polish).

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

[18] Strobin, L., *Wyszukiwanie zależności semantycznych w grafowych bazach danych z zastosowaniem logiki rozmytej i algorytmów genetycznych*, Ph.D. thesis, Politechnika Łódzka, 2017.

[19] Superson, I. and Niewiadomski, A., *Pozyskiwanie wiedzy z relacyjnych baz danych: wielopodmiotowe podsumowania lingwistyczne*, Economic Studies. University of Economics in Katowice, 2014, pp. 301–304, (in Polish).