INFORMATICA, 2020, Vol. 31, No. 4, 841–856

DOI: https://doi.org/10.15388/20-INFOR428

Linguistic Summaries in Evaluating Elementary Conditions, Summarizing Data and Managing Nested Queries

Pavol SOJKA[1], Miroslav HUDEC[1,2,*], Miloš ŠVAŇA[2]

[1] *Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1, Bratislava, Slovakia*
[2] *Faculty of Economics, VSB – Technical University of Ostrava, Sokolská třída 33, Ostrava, Czech Republic*
e-mail: miroslav.hudec@euba.sk

Received: January 2020; accepted: August 2020

Abstract. Data users are generally interested in two types of aggregated information: summarization of the selected attribute(s) for all considered entities, and retrieval and evaluation of entities by the requirements posed on the relevant attributes. Less statistically literate users (e.g. domain experts) and the business intelligence strategic dashboards can benefit from the linguistic summarization, i.e. a summary like *the most of customers are middle–aged* can be understood immediately. Evaluation of the mandatory and optional requirements of the structure P_1 *and most of the other posed predicates should be satisfied* is beneficial for analytical business intelligence dashboards and search engines in general. This work formalizes the integration of aforementioned quantified summaries and quantified evaluation into the concept of database queries to empower their flexibility by, e.g. the nested quantified query conditions on hierarchical data structures. Next, this approach contributes to the mitigation of the empty answer problem in data retrieval tasks. Thus, the strategic and analytical dashboards as well as query engines might benefit from the proposed approach. Finally, the obtained results are illustrated on examples, the internal and external trustworthiness is elaborated, and the future research topics and applicability are discussed.

1. Introduction

Databases usually contain a large number of entities and their attributes. Formally, a database can be expressed as a set of pairs (Skowron *et al.*, 2015)

$$T_p = (U_p, A_p), \tag{1}$$

where U_p is an universe of entities (records) and A_p is a set of attributes in a table Tp, $p = 1, \dots, n$. In such tables, rows are labelled by entities and columns by attributes.

[*] Corresponding author.

Generally, data users are interested in two types of queries, which might be expressed as vertical and horizontal aggregations. In the former, statistical functions, such as means, deviations and distributions are used to explain the entities' attributes considering the set of entities U (1), e.g. *the average altitude of all municipalities is m, whereas the standard deviation is s*, where $m, s \in \mathbb{R}$. In the latter, users are interested in entities that best satisfy the compound predicate, i.e. in finding the best entities considering the predicates posed on a subset of attributes A (1), such as *altitude BETWEEN* (1000, 1200) **and** *pollution* $\leqslant$50 **and** *population density* $<$100 **and** *number of sunny days* $\geqslant$120 **and** *percentage of arable land* $\geqslant$20. In such queries, users have to express requirements by numbers, even though uncertainties about the borderline cases might appear (Keefe, 2000). The third case, nested sub-queries posed against the 1 : N relationships (e.g. between the relations *district* and *municipality*) may require merging horizontal and vertical aggregations, such as *select districts, where unemployment rate* $\geqslant$30 **and** *percentage of respiratory diseases* $>$25 **and** *the average pollution in their respective municipalities is higher than the limit value*. For the data users, the natural way to express the requirements is by linguistic terms. The same holds for interpreting the summaries where, despite the broad use of statistical functions, they are suitable for domain experts having a certain level of statistical literacy (Hudec *et al.*, 2018).

The literature has already recognized the limitations of the classical or two-valued approaches and provided solutions for the various situations. The vertical aggregation has been empowered by the so-called linguistic summaries initially proposed by Yager (1982) and emphasized (Yager *et al.*, 1990) that summaries should not be as terse as means. Since then, the theory of linguistically summarized sentences has been extensively researched by many scholars and applied in a variety of fields. A detailed, although not a very recent review, can be found in Boran *et al.* (2016). Less statistically literate users (e.g. domain experts and the general public) can benefit from such a summarization (Hudec *et al.*, 2018; Schield, 2011). Through this approach, we are able to provide an overall overview of one attribute or relations among several attributes in a dataset, such as *about half of the municipalities have the population density around the mean value*, or *the majority of young customers buy items in late evenings*. Such summaries might improve the informativeness of business intelligence strategic dashboards, for instance.

A query against the data stored in a database provides a formal description of the entities of interest to the user posing this query (Hudec and Vučetić, 2015; Kacprzyk *et al.*, 2000). Limitation of the two-valued logic in the database query conditions has been mitigated by the fuzzy query approaches like (Bosc and Pivert, 1995; Hudec, 2009; Kacprzyk and Zadrożny, 1995; Wang *et al.*, 2007). In this way, the most relevant entities with respect to user needs are retrieved together with their matching degrees, i.e. the closeness to the full satisfaction. An example of such query is *select customers having a high number of orders and low payment delay*. Next, a user might be interested in entities that meet the majority of requirements. Such a query is of the structure *the most of atomic requirements* $\{P_1 \ldots P_n\}$ *should be met* (Kacprzyk and Ziółkowski, 1986). However, this approach is not able to make distinction between the mandatory and optional requirements. Further, linguistic summaries have shown their applicability as nested subqueries in the hierarchical data structures, e.g. *select regions where the most of municipalities meets the requirement P* (Hudec, 2016). More complex nested queries require the integration of the vertical and horizontal aggregations.

The foundation for all the aforementioned approaches is the theory of fuzzy sets introduced by Zadeh (1965), the theory of fuzzy logic based on the theories of many-valued logics and fuzzy sets, and the theory of aggregation functions summarized in Dubois and Prade (2004), Beliakov *et al.* (2007). Thus, the methodology of our work is based on the key findings in these fields.

The research questions in this work are the following: the problem of merging the horizontal and vertical aggregation and the formalization of mandatory and optional predicates in quantified queries, and a subsequent proposal of a suitable integration. By this approach, we can cover the gap in the merging of quantified summarization with evaluation. In addition, when a conjunctively expressed query condition consists of a larger number of predicates, an empty answer might appear. The proposed aggregation by the fuzzy quantifier *most of* is a semantically different contribution than the existing approaches covering the empty answer problem (Bosc *et al.*, 2009, 2008, 2007; Smits *et al.*, 2014) and therefore it augments the established ones.

The remainder of the paper is organized as follows: Section 2 provides a brief explanation of the main aspects of linguistic summaries, which is necessary for the subsequent sections. Section 3 is dedicated to formalizing the quantified evaluations of entities and aggregating them with quantified summaries, whereas Section 4 demonstrates the results on illustrative situations, evaluates the validity of results, discusses research questions, raises future research topics and applicability. Finally, Section 5 answers research questions and concludes the paper.

2. Linguistic Summaries in Brief

This section studies the relevant theoretical aspects of data summaries by short quantified sentences of natural language. A basic structure of such sentence has the form *Q entities in a dataset are P* where *Q* is a linguistic quantifier such as *most of*, *about half* and *few*, and *P* is an elementary or compound predicate. The truth value (or validity) is calculated in the following way (Yager, 1982)

$$v\big(Qx\big(P(x)\big)\big) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_P(x_i)\right), \tag{2}$$

where n is the number of entities or the scalar cardinality of a dataset (a universe of entities U_p (1)), $y = \frac{1}{n}\sum_{i=1}^{n}\mu_P(x_i)$ is the proportion of entities in a dataset that satisfy predicate P, $\mu_P(x_i)$ is the matching degree of entity x_i to predicate P, and μ_Q is the membership function of a chosen relative quantifier. The truth value v assumes values from the unit interval.

Formalization of fuzzy relative quantifiers can be carried out by using three methods: sigma-counts (Zadeh, 1983), Ordered Weighted Averaging (OWA) operator (Yager, 1988)

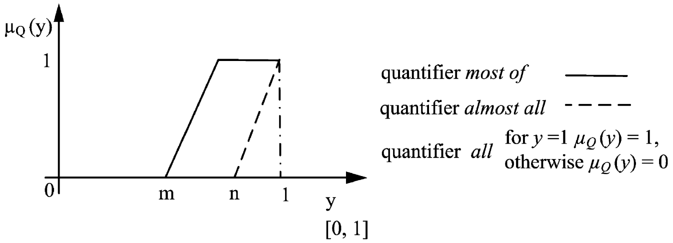

Fig. 1. Parametrized linguistic quantifier *most of*, where y is the proportion of entities that meet the predicate P (see, Eq. (2)).

and Competitive Type Aggregation (Yager, 1984). The sigma-count method is adopted for this work, because it allows the quantifiers and predicates to be modelled in the same way, which simplifies the applicability, and therefore is more intuitive for diverse users. Within this method, the quantifier *most of* is formalized by an increasing (usually linear) function. It can be constructed independently by equations offered in Kacprzyk and Zadrożny (2005) or as one granule from the family of uniformly distributed relative quantifiers constructed on the [0, 1] interval (Hudec, 2016). When expressed by parameters, the quantifier *most of* yields (see, Fig. 1):

$$\mu_Q(y) = \begin{cases} 1, & \text{for } y \geqslant n, \\ \frac{y-m}{n-m}, & \text{for } m < y < n, \\ 0, & \text{for } y \leqslant m, \end{cases} \tag{3}$$

where $0.5 \leqslant m \leqslant n \leqslant 1$. When $m = n = 1$, the quantifier becomes the crisp quantifier *all*, whereas when, e.g. $0.8 \leqslant m < n = 1$, the quantifier expresses the term *almost all*.

Analogously, the quantifier *about half* can be expressed by a symmetric triangular fuzzy number centred around the value of 0.5 ($\mu_Q(0.5) = 1$). The quantifier *few* is expressed by a decreasing function ($\mu_Q(0) = 1$, $\mu_Q(1) = 0$).

The linguistic terms *low*, *medium*, *around m* and *high* can be formalized by an L fuzzy set, a trapezoidal fuzzy set, a triangular fuzzy set and a linear gamma fuzzy set, respectively, as illustrated in Fig. 2.

Generally, fuzzy sets can be formalized by non-linear functions. In this work, we adopted the linear ones due to their simplicity for the end users. We used the same adoption also for the relative quantifiers.

In this work, we apply the basic structure of linguistic summaries (2), evaluations expressed by the quantifier *most of* Kacprzyk and Zadrożny (2005), that is further parametrized in Hudec (2016) and aggregation functions (Beliakov *et al.*, 2007), in order to explore the raised research questions. A review of the other types of linguistic summaries can be found in Lesot *et al.* (2016), whereas a review of applicability can be found in Boran *et al.* (2016). The solution of a summary is the validity or truth value of the evaluated quantified sentence, not a set of retrieved entities from (1).

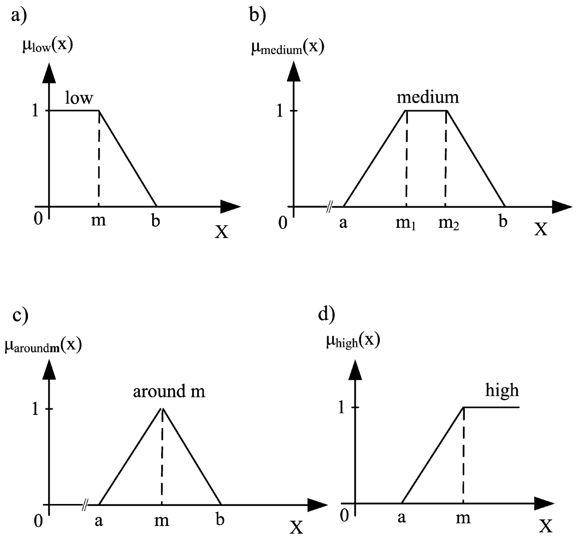

Fig. 2. Fuzzy sets: a) L fuzzy set, b) trapezoidal, c) triangular, d) linear gamma.

3. Evaluation of Atomic Conditions by Quantified Summaries

This section studies linguistic summaries employed as aggregations of elementary requirements in the evaluation of entities.

3.1. *Evaluation of Optional Atomic Conditions*

In database queries, the usual way of selecting the relevant entities is realized via the conditions expressed by conjunction (*AND* operator) and disjunction (*OR* operator). The former cannot cover the aggregation of mandatory and optional requirements, because all the requirements are mandatory. If only one atomic condition from a larger set is rejected, the overall matching degree is zero (zero is the absorbing element in conjunction). The latter is based on the substitutability principle, i.e. one satisfied atomic requirement is sufficient.

Let us recall the standard classification of aggregation functions (Dubois and Prade, 2004). Conjunctive aggregation functions are characterized by $A(\mathbf{x}) \leqslant \min(\mathbf{x})$, disjunctive by $A(\mathbf{x}) \geqslant \max(\mathbf{x})$, averaging by $\min(\mathbf{x}) \leqslant A(\mathbf{x}) \leqslant \max(\mathbf{x})$, and remaining aggregation functions are called mixed, where $\mathbf{x}$ is a vector, $\mathbf{x} = (x_1, \dots, x_n)$.

Thus, the following problems in conjunctive aggregation might appear. First, the presence of mandatory and optional requirements, which was addressed by the asymmetric *AND IF POSSIBLE* conjunction (Dujmović, 1975; Bosc and Pivert, 2012) and axiomatized in Hudec and Mesiar (2020). Second, the aforementioned *empty answer problem*, i.e. cases when not a single record meets a larger set of atomic conditions. Third, all atomic predicates might be optional where the principle *the more the better* holds, i.e. an entity is preferred over another one if it satisfies more predicates. Thus, to exclude weakly per-

Table 1
Example of the quantified evaluation of optional conditions.

Entity	P_1	P_2	P_3	P_4	P_5	P_6	P_7	Solution (4)*	Arith. mean
e1	0.8	0.9	0.6	0	1	0.7	0.8	0.53	0.69
e2	0.1	0.4	0.1	0.3	0.2	0.1	0.1	0	0.19
e3	1	0	1	0.95	1	1	0.8	0.92	0.82
e4	0.3	0	0.4	0.2	1	0.7	0.85	0	0.49
e5	0.8	0.8	0.85	0.9	0.9	0.9	0.8	1	0.85
e6	0	0	0	0	0	0	0	0	0
e7	1	1	1	1	1	1	1	1	1

*Where $m = 0.5$ and $n = 0.85$ in (3).

forming entities, and on the other side to mitigate the empty answer problem, conjunction is relaxed by the quantifier *most of*, that is, *most of atomic predicates should be satisfied*.

The query relaxation by the fuzzy relative quantifier: *the most of atomic conditions should be satisfied* is initially suggested by Kacprzyk and Ziółkowski (1986). It is formalized by the quantified summaries (2), (3) as follows:

$$v(x) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_{P_i}(x)\right), \tag{4}$$

where n is the number of atomic predicates posed on a subset of attributes A (1), $y = \frac{1}{n}\sum_{i=1}^{n}\mu_{P_i}(x)$ is the proportion of atomic predicates P_i that are satisfied by the entity x being evaluated and μ_Q is the formalization of the quantifier *most of*. The truth value v assumes values from the unit interval.

This approach deals with the presumption that all atomic conditions are optional, but the majority of them should be met. When all atomic conditions are met very weakly, then due to the quantified evaluation aggregated by the quantifier (3) the solution is zero. Hence, the aggregation for low values behaves like a nilpotent conjunction (e.g. Łukasiewicz t-norm $T_l(\mathbf{x}) = \max(0, \sum_{i=1}^{n} x_i - (n-1))$, where for instance $T_l(0.2, 0.1, 0.2, 0.3) = 0)$ and for high values behaves like nilpotent disjunction (e.g. Łukasiewicz t-conorm $S_l(\mathbf{x}) = \min(0, \sum_{i=1}^{n} x_i)$, where for instance $S_l(0.9, 0.8, 0.9, 0.8) = 1)$. The disjunctive nilpotent observation holds only when the majority of atomic conditions is significantly met, e.g. $y \geqslant n$ (see, Fig. 1 and (4)). Thus, value 1 is not automatically the absorbing element when several elementary conditions are met. The result for $(0.2, 0.8, 0.4, 1)$ is 0.29 ($m = 0.5$, $n = 0.85$). Similarly, for $(0, 0.5, 0.9, 0.9)$ the solution is 0.21. In these cases, the behaviour is averaging. Clearly, this aggregation is a mixed one. These observations can be further axiomatized into the frame of the standard classification of aggregation functions, which is a future research topic.

This aggregation is illustrated in Table 1, where the entity *e2* is rejected, even though it meets weakly all the atomic conditions (the property of conjunctive nilpotency). Entities *e1* and *e3* fail to meet one of the atomic requirements, but are evaluated as two acceptable ones because they meet significantly all the other requirements, with *e3* being almost the ideal solution. These cases do not behave like disjunctive ones, because few fully satisfied predicates do not guarantee straightforwardly the solution equal to 1. Finally, entity *e5* is evaluated as the ideal one due to very significant satisfaction of all requirements, even though not a single requirement is fully met (the property of disjunctive nilpotency, which appears only when most of the predicates are met significantly). For completeness, a solution by the arithmetic mean is shown in the last column. The arithmetic mean is not able to exclude entities *e2* and *e4*. Clearly, by (4) all requirements are considered as optional and compensative when the proportion of (fully or partially) satisfied elementary requirements is high, usually greater than 0.5, for covering the natural meaning of terms *majority* or *most of*. Entities *e6* and *e7* show that the boundary conditions of aggregation functions are met.

3.2. *Evaluation of Mandatory and Optional Atomic Conditions*

In many cases, several atomic requirements are mandatory, while the other ones are optional, and moreover, if a higher proportion of optional requirements is satisfied, then the entity is more suitable. In order to cover this aggregation requirement, we have modified the Eq. (4) in the following way

$$v(x) = \left(\bigwedge_{i=1}^{r} P_i(x)\right) \wedge \mu_Q\left(\frac{1}{s}\sum_{j=1}^{s} \mu_{P_j}(x)\right), \tag{5}$$

where r is the number of mandatory requirements and s is the number of optional requirements (usually $r \ll s$), and x is an evaluated entity.

A suitable method for formalizing *AND* connective (conjunction) in fuzzy logic is by triangular norms (or in short t-norms), because of the desirable properties (monotonicity, associativity, symmetry and the presence of a neutral element) (Hájek, 1998). The four basic t-norms are (Klement *et al.*, 2005): minimum t-norm, product t-norm, Łukasiewicz t-norm and drastic product. The least suitable is drastic product due to its very restrictive nature and non-continuity. The product t-norm and Łukasiewicz t-norm have a downward reinforcement property, whereas the minimum t-norm has the property of idempotency (Beliakov *et al.*, 2007). For instance, when using the Łukasiewicz t-norm, the solution is greater than zero when both mandatory and quantified parts are significantly satisfied. We have two conjunctions in Eq. (5): among mandatory and between mandatory and quantified requirements. In addition, there exist conjunctive functions which do not meet all the axioms of t-norms, e.g. $C(a, b) = a^v \cdot b^w$ (Beliakov *et al.*, 2007; Hudec and Vučetić, 2019), where $v > 1$ and $w > 1$ indicate the importance of predicates. Observe that for $v = w = 1$ we get the product t-norm and for $v < 1$, $w < 1$ and $v + w = 1$ we get the geometric mean. Although the other functions could be examined, this work is focused on the minimum t-norm.

An illustrative example is searching for suitable accommodation units. The atomic requirements can be low price, high safety rating, altitude above sea level around 1500 meters, low pollution, small population density, short distance to the nearest grocery store, high rating of the accommodation unit and the like. It is highly presumable that none of the evaluated units meets all predicates in a pure conjunctive way. However, the user might express that the low price and high safety ratings are mandatory requirements, whereas the others are optional.

3.3. *Evaluation of Nested Queries*

This class of queries is convenient for the 1 : N relationships in a database such as district-municipality and customer-invoice. An example of a nested query condition might be as follows: *select districts where the most of municipalities have high pollution and high number of respiratory diseases*. The answer should be a list of districts which fully or partially meet the condition ranked downward from the best by the intensity of matching degrees.

The procedure for calculating validities is created straightforwardly as the extension of (2) in the following way (Hudec, 2016):

$$v_j(x_j) = \mu_Q\left(\frac{1}{N_j}\sum_{i=1}^{N_j}\mu_P(x_{ij})\right), \quad j = 1\dots C, \qquad \sum_{j=1}^{C} N_j = n, \tag{6}$$

where n is the number of entities in the entire database, N_j is the number of entities in the group j (in our case, municipalities belonging to the district j), C is the number of groups in the database (in our case, districts), v_j is the validity of the summary for the j-th group, and $\frac{1}{N_j}\sum_{i=1}^{N_j}\mu_P(x_{ij})$ is the proportion of entities x_i in the j-th group that satisfy the quantified query condition. Thus, this query should be executed C times, i.e. to calculate the validity for each unit on the "1" side in the 1 : N relationship. Generally, P can be an elementary or a compound predicate, i.e. several atomic conditions merged by a logical connective. In the aforementioned illustration, we have two elementary conditions (high pollution and high number of respiratory diseases).

This method integrates the vertical aggregation explained in Section 2 and the horizontal one discussed in Section 3.1. Further, we may have a quantified condition on a higher hierarchical level and a quantified summary on a lower level, e.g. *select districts where the most of* $\{P_1 \dots P_n\}$ *are met and most of the respective municipalities have high value of* R_g, where $P_1 \dots P_n$ are atomic requirements posed on districts'attributes, whereas R_g is a requirement related to the municipalities' attribute.

In the case of aggregating the conjunction of atomic predicates on the "1" side and the quantified condition on the "N" side, the formalization is as follows:

$$v(x_j) = \left(\bigwedge_{r=1}^{R} P_r(x_j)\right) \wedge \mu_Q\left(\frac{1}{N_j}\sum_{i=1}^{N_j}\mu_P(x_{ij})\right), \quad j = 1\dots C, \qquad \sum_{j=1}^{C} N_j = n, \tag{7}$$

where the parameters have the same meanings as in the respective equations, x_j is an entity in a table on the side "1", and x_{ij} is a record in the related table on the side "N" categorized under the record x_j.

Table 2
Example of the aggregation of the mandatory and quantified evaluation of optional conditions by conjunction expressed as minimum t-norm and the quantifier *most of* formalized by $m = 0.5$ and $n = 0.85$ (see Fig. 1).

Unit	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	Solution (5)
u1	1	0.8	0.9	0.2	0	0.9	0.4	0.85	0.9	0.6	0.268
u2	1	0	1	1	1	0.75	0.8	0.9	0.6	0.9	0
u3	0.75	0.85	0.9	0.85	0.95	0.7	0	1	0.9	0.85	0.75
u4	1	0.9	0.2	0.3	0	0.15	0.85	0.35	0.7	1	0
u5	0	0.3	1	1	1	0.25	0.75	0.3	0.75	0.8	0
u6	1	0.8	0.9	0.85	0.95	0	1	1	0.9	0.65	0.786
u7	0.2	0.8	0.3	0.35	0.8	0.35	0	0.4	0.2	0.5	0
u8	0	0.8	0.95	0.75	0.9	0.45	1	0.55	0	0.65	0
u9	0.3	0.8	0.75	0.95	0.9	0	0.85	0.85	0.75	0.8	0.3
u10	1	0.2	0.9	0	0.5	0.65	0.4	0.8	0.35	0.55	0.054
u11	0	0	1	1	0.9	0.25	0.85	0.9	0.55	0	0
u12	1	1	0	0	0.2	0.15	0.2	0.6	0.5	1	0

The right part of (7) aggregates on the set of entities U_2 (1), while the left part aggregates on the set of attributes A_1.

The next section is devoted to examples and to a discussion illustrating the suggested approach.

4. Illustrative Examples and Discussion

This section illustrates the approach proposed in Section 3 on three examples, provides a discussion and a brief reflection upon the possible applicability.

4.1. *Illustrative Examples*

This part consists of the following three examples: (i) evaluation of the mandatory and optional atomic conditions, (ii) summarization of the particular attributes for all records and (iii) nested query for hierarchical data.

An example of aggregating the mandatory and optional atomic conditions

Hypothetical twelve accommodation units and matching degrees of ten atomic requirements relevant for a hypothetical guest are shown in Table 2. In the example, low price (P_1) and high safety rating (P_2) are mandatory requirements, whereas the majority of the remaining ones: altitude above the sea level around 1500 meters (P_3), low pollution (P_4), small population density (P_5), high rating of the accommodation unit (P_6), short distance to the nearest grocery store (P_7), high number of sunny days (P_8), short distance to the funicular (P_9) and high cleanness (P_{10}), should be met. The minimum t-norm was adopted for both conjunctions in (5), among the mandatory requirements and the aggregation with the quantified part.

When we consider all the requirements to be mandatory, then not a single unit is selected, since we can find a non-satisfied elementary predicate for all the units. The probability of the empty answer problem increases when the number of atomic requirements increase, as is the case in this example. Thus, the user either abandons the holiday (a less probable alternative), or relaxes the requirement. It can usually be realized by reducing the number of atomic conditions or relaxing some of the atomic predicates (e.g. for the predicate *short distance* the user will accept a bit longer distance). The option suggested in this work is the quantified relaxation. In the relaxed quantified condition, if the price or the safety is not satisfied, then the unit is rejected regardless of the degrees of satisfaction of the other requirements. A non-satisfied requirement from the set $\{P_3, \dots, P_{10}\}$ is not a reason for rejection. In Table 2 we can see that the most suitable unit is *u6*, followed by *u3*. Although not a single unit fully satisfies all the expectations, two units are very close to meeting them, and therefore are the most preferable, although not the ideal alternatives.

An example of summarized information

Let us have the same data as in Table 2. A hypothetical manager in a tourism agency wishes to know whether the most of accommodation units have a high cleanness rate, and also whether the most of accommodation units have a short distance to the nearest grocery store. In this example, we have two summarization tasks related to one attribute each. Focus in the calculations is on the predicates *high cleanness, P10* and *short distance, P7*, respectively, where the matching degrees are shown in respective columns.

By applying Eq. (2), where $n = 12$ (number of entities) and parameters of the quantifier *most of* (Fig. 1) are 0.5 and 0.85 we obtain the validity of the sentence *the most of accommodation units have a high cleanness rate* equal to 0.94. The high truth value indicates that this sentence is accepted. However, the validity of the second sentence is 0 (the proportion $\frac{1}{n}\sum_{i=1}^{n}\mu_P(x_i)$ is equal to 0.46), and therefore this sentence is rejected.

Moreover, linguistic summaries are able to offer an alternative answer when the initial quantified sentence is of an insufficient validity (Hudec *et al.*, 2018). The proportion of the afore evaluated second sentence (0.46) indicates that the most suitable quantifier is *about half*, and therefore the answer is not only that the validity of the initial sentence is zero, but we can provide an alternative summary: *about half of accommodation units have short distance to the nearest grocery store*. Similarly, the user can examine the remaining predicates of interest.

An example of a nested query

An analyst is interested in revealing highly attractive districts where the most of accommodation units have a low rating, i.e. districts, which might have a significant increase in the number of visits if the accommodation units improved their quality.

Let us have hypothetical accommodation units situated in four districts as is shown in Tables 3 and 4, where we have already calculated the matching degrees to atomic predicates. For instance, the attribute rating assumes values from the [0, 100] interval. The predicate *low rating* is formalized by a L fuzzy set (see Fig. 2) with parameters $m = 10$ and $b = 30$. Accommodation unit *u4* has an aggregated rating 13 and therefore belongs to the predicate *low rating* with the degree of 0.85. The matching degrees of the other units are calculated in the same way.

The analyst sees that the most problematic district is *d1*, followed by *d3*, whereas *d2* meets this requirement very weakly and *d4* is not a problematic district. So, the analyst

Table 3
The evaluated districts by their own predicates and quantified predicates of their respective accommodation units (see Table 4).

District	High attractiveness*	Most of**	Solution (7)
d1	0.90	0.6607	0.6607
d2	0.85	0.1111	0.1111
d3	0.45	1	0.45
d4	0	0.9524	0

* By, e.g. an index of attractiveness provided by an agency.
** The solution of the requirement *the most of accommodation units have a low rating*.

Table 4
Accommodation units' matching degrees to predicates.

Unit	Long distance to bus stop	Low rating	Low cleanness	District
u1	0.4	0.9	0.6	d1
u2	0.8	0.75	0.9	d1
u3	0	0.7	0.85	d1
u4	0.85	0.85	1	d1
u5	0.75	0.25	0.8	d1
u6	1	0.9	0.65	d1
u7	0	0.85	0.5	d1
u8	1	0.65	0.65	d1
u9	0.85	0	0.8	d2
u10	0.4	0.65	0.55	d2
u11	0.85	0.3	0	d2
u12	0.2	0.15	1	d2
u13	0.9	0.4	0.25	d2
u14	1	0.45	0.9	d2
u15	1	0.9	0.75	d2
u16	0.85	1	0	d2
u17	1	1	0.35	d2
u18	0.25	0.9	0	d3
u19	0	0.85	0.45	d3
u20	0	0.75	0	d3
u21	0.65	0.95	0.55	d3
u22	0.5	1	0.35	d3
u23	0.15	1	0.75	d3
u24	0	0.85	0.65	d3
u25	0.85	0.95	0.65	d4
u26	0.15	1	0.35	d4
u27	0.25	0.45	0	d4
u28	0.95	0.85	0.35	d4
u29	0.1	0.75	0.55	d4
u30	1	1	0.35	d4

has recognized the districts where the improvements in the quality of the accommodation units might bring a significant increase in the number of visits.

Similarly, the analyst can evaluate the other attributes of accommodation units such as distances to relevant points and prices, as well as the attributes of the districts, where the accommodation units are situated, such as pollution and safety.

4.2. *Discussion*

The proposed approach has a high applicability potential. Quantification of optional requirements and their aggregation by the conjunction with the mandatory ones augments the database querying possibilities by providing further ways of expressing the users' requirements in diverse tasks and mitigates the empty answer problem. This holds especially for the cases, when the users pose a higher number of atomic requirements, but are not sure which of them might be excluded in the case of an empty answer, or they are interested in seeing which entities satisfy at least the majority of optional requirements.

For these tasks, we adopted the equation for the calculation of the validities of quantified sentences (2) in such a way that instead of the proportion of entities that meet the summarizer, we have the proportion of atomic requirements met by the evaluated entities (4). Further, we aggregated this equation with the mandatory requirements by a conjunctive function (5). In this work, we adopted the minimum t-norm for conjunction. The future research work should be focused on examining the suitability of the other conjunctive functions between the mandatory and quantified part, and among the mandatory predicates.

The initial equation for linguistic summaries (2) remains suitable for expressing knowledge about the attribute in the entire data set. Less statistically literate users (e.g. domain experts on the business intelligence strategic dashboards, and general public on the official statistics data dissemination websites (Hudec *et al.*, 2018; Schield, 2011) or the eInforming stage of smart cities (Terán *et al.*, 2016)) might benefit especially from the summarization by short quantified sentences.

Business intelligence visualization considers strategic, operational and analytical dashboards (Vaisman and Zimányi, 2014). Strategic dashboard is a collection of multiple visual components (e.g. charts and key performance indicators) on a single view so the main messages can be monitored at a glance (Few, 2006). Hence, the main goal is to explain what is going on, not to explain the reasons for such a behaviour. Therefore, the linguistic summaries in their initial sense (2) are suitable for strategic dashboards. They are able to express the model of behaviour linguistically, and are especially suitable for the top-managers. Further, a pattern such as *the most of visits from French speaking countries is in spring months* cannot be easily interpreted graphically. As we see, the solutions of linguistically summarized sentences do not explain the reasons for such behaviour, but indicate what is going on and emphasize the most perspective or the most problematic cases.

Evaluation of entities by quantified aggregation (5) and (7) is suitable for analytical dashboards to support the tasks focusing on the recognition of the perspective or problematic entities by the conditions expressed via the natural language expressions in order to cover the users' uncertainties regarding the conditions expressed by a larger number of requirements.

Users searching for the most suitable entities, like the accommodation units in the above-mentioned illustrative examples, and for the summarized information regarding the various aspects of our society (e.g. key statistical figures released by the national statistical institutes) might benefit from this approach. A promising application field is empowering the concept of smart cities by flexible summarizing of the developments in the city and by supporting the search for the most suitable entities such as dwelling units, the most acceptable locations and similar participants for the discussion groups. The next task for the future work will be the development of intuitive user-friendly interfaces.

We have used the sigma-counts method within the *computing with words* methodology for the construction of fuzzy predicates and relative quantifiers. Adjectives such as *high* and quantifiers such as *most of* are always expressed by increasing functions, regardless of their translation to other human languages. The same holds for the adjective *low*, which is always expressed by a decreasing function. We have also used methods for aggregating entities and attributes into compound user requirements.

The data are stored in databases as numbers that in many cases pretend a precision, since the real data are frequently not available as precise numbers, but they are more or less non-precise or fuzzy. The internal trustworthiness shows that the results are less sensitive to the factors such as a small imprecision in the data and the user's linguistic requirements formalized as fuzzy sets. Fuzzy logic queries allow similar entities to be similarly evaluated, or to similarly contribute to the summaries. Next, the boundary conditions, the key requirements in the quantified aggregation, are satisfied (see Table 1, entities *e6* and *e7*). The boundary condition is by definition satisfied for the linguistic summary (2) and the conjunction, and therefore it holds for (5) and (7). And the monotonicity (if the matching degree of one atomic predicate increases, whereas the other remains the same, the solution remains the same or increases) is a matter of direct verification.

External trustworthiness generalizes the other situations and data sets. If the user query contains a large number of elementary predicates for evaluation or a large number of entities for summarization, the limitation is the computational capacity. Mathematically, aggregation is not limited. When we consider, for instance, business intelligence dashboards, the set of attributes for evaluation is not large. Regarding the summarization on an increasing number of records, for example, in the fact tables in data warehouses, the proportion in (2), or the quantified part in (7) is a distributive operation (in fact, the addition is, and therefore, previous sum is used). So, we can build results on the previously calculated proportions.

Next, data incompleteness also occurs in databases (the value is simply unknown or inconvenient for the provider to offer). It influences the conjunction in the same way as in the classical SQL. In the quantified condition, it is not so problematic, because when we replace the missing value by the matching degree 0, we do not automatically exclude any records. Adjustment of the missingness-tolerant evaluation suggested for the logic scores of preferences (Dujmović, 2018) is a topic for the future research. Generally, we can assign values from 0 (full penalty for missingness) to 1 (full tolerance). This issue is more relevant to the conjunctive and quantified evaluation of attributes than to the summarization of a large number or records.

5. Conclusions

In our data intense society, we face the problems of diverse needs for the aggregation of atomic requirements for the evaluation of entities and for the explanation of summarized information. In order to contribute, we raised the research question of merging the quantified evaluation and the quantified summarization, as well as aggregating the mandatory and optional quantified predicates in a fuzzy environment, where the requirements are expressed via fuzzy sets.

The answer is that the quantified evaluation (so-called horizontal quantified aggregation) and the data summarization (vertical aggregation) are supported by the fuzzy relative quantifiers and therefore can be straightforwardly merged to answer complex queries on $1:\mathrm{N}$ relationships such as *select regions where most of atomic predicates posed on regions are satisfied and most of their respective municipalities have a high value of the attribute* A. In this work, the relative quantifiers are formalized by sigma-counts, and therefore, the quantifiers and predicates are modelled by the same method, which simplifies the applicability and is more intuitive for users.

For the quantified evaluation, we should consider cases where some of the atomic requirements are mandatory. In this case, we should aggregate mandatory and optional quantified requirements by a conjunctive function. The conjunction in the present work is realized by the minimum t-norm. For the future research, we shall consider other conjunctive functions to cover various conjunctive aspects of the aggregation of mandatory requirements and optional quantified requirements. The quantified aggregation of atomic predicates is a relaxation of conjunction, which augments the existing approaches to the mitigation of empty answer problems.

The suggested approach is demonstrated on examples in order to illustrate the diverse needs of the users. Real-world tasks such as flexible recommendations, informing and searching in smart cities and business intelligence dashboards might benefit from the results of this work.

Funding

This paper was supported by the SGS project No. SP2020/125 of the Ministry of Education, Youth and Sports of the Czech Republic. Also the support of the project VEGA No. 1/0373/18 by the Ministry of Education, Science, Research and Sport of the Slovak Republic is kindly announced.

References

Beliakov, G., Pradera, A., Calvo Sánchez, T. (2007). *Aggregation Functions: A Guide for Practitioners*. Springer-Verlag, Berlin, Heidelberg.

Boran, F.E., Akay, D., Yager, R.R. (2016). An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications*, 61, 356–377.

Bosc, P., Pivert, O. (1995). SQLf: a relational database language for fuzzy querying. *IEEE Transactions on Fuzzy Systems*, 3, 1–17.

Bosc, P., Pivert, O. (2012). On four noncommutative fuzzy connectives and their axiomatization. *Fuzzy Sets and Systems*, 202, 42–60.

Bosc, P., Hadjali, A., Pivert, O. (2007). Weakening of fuzzy relational queries: and absolute proximity relation-based approach. *Mathware and Soft Computing*, 14, 35–55.

Bosc, P., Hadjali, A., Pivert, O. (2008). Empty versus overabundant answers to flexible relational queries. *Fuzzy Sets and Systems*, 159, 1450–1467.

Bosc, P., Brando, C., Hadjali, A., Jaudoin, H., Pivert, O. (2009). Semantic proximity between queries and the empty answer problem. In: *Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference*, Lisbon, pp. 259–264.

Dubois, D., Prade, H. (2004). On the use of aggregation operations in information fusion processes. *Fuzzy Sets and Systems*, 142, 143–161.

Dujmović, J. (1975). Extended continuous logic and the theory of complex criteria. In: *Series Mathematics and Physics*, Vol. 498–541. Univ. Beograd. Publ. Elektrotechn. Fak., Belgrade, pp. 197–216.

Dujmović, J. (2018). *Soft Computing Evaluation Logic: The LSP Decision Method and Its Applications*. Wiley, IEEE Computer Society, Hoboken.

Few, S. (2006). *Information Dashboard Design: The Effective Visual Communication of Data*. O'Reilly Media.

Hájek, P. (1998). *Meta Mathematics of Fuzzy Logic*. Kluwer Academic Publishers, Dordrecht.

Hudec, M. (2009). An approach to fuzzy database querying, analysis and realisation. *Computer Science and Information Systems*, 6, 127–140.

Hudec, M. (2016). *Fuzziness in Information Systems – How to Deal with Crisp and Fuzzy Data in Selection, Classification, and Summarization*. Springer, Cham.

Hudec, M., Mesiar, R. (2020). The axiomatization of asymmetric disjunction and conjunction. *Information Fusion*, 53, 165–173.

Hudec, M., Vučetić, M. (2015). Some issues of fuzzy querying in relational databases. *Kybern*, 51, 994–1022.

Hudec, M., Vučetić, M. (2019). Aggregation of fuzzy conformances. In: *10th International Summer School on Aggregation Operators*, AGOP 2019, Olomouc, Czech Republic.

Hudec, M., Bednárová, E., Holzinger, A. (2018). Augmenting statistical data dissemination by short quantified sentences of natural language. *Journal of Official Statistics*, 34, 981–1010.

Kacprzyk, J., Zadrożny, S. (1995). Fquery for access: fuzzy querying for windows-based DBMS. In: Bosc, P., Kacprzyk, J. (Eds.), *Fuzziness in Database Management Systems*. Physica-Verlag, Heidelberg, pp. 415–433.

Kacprzyk, J., Zadrożny, S. (2005). Protoforms of linguistic database summaries as a human consistent tool for using natural language in data mining. *International Journal of Software Science and Computational Intelligence*, 1, 1–11.

Kacprzyk, J., Ziółkowski, A. (1986). Database queries with fuzzy linguistic quantifiers. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-16(3), 474–479.

Kacprzyk, J., Pasi, G., Vojtáš, P., Zadrożny, S. (2000). Fuzzy querying: issues and perspectives. *Kybernetika*, 36, 605–616.

Keefe, R. (2000). *Theories of Vagueness*. Cambridge University Press, Cambridge.

Klement, E.P., Mesiar, R., Pap, E. (2005). Triangular norms: basic notions and properties. In: Klement, E.P., Mesiar, R. (Eds.), *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*. Elsevier, Amsterdam, pp. 17–60.

Lesot, M.-J., Moyse, G., Bouchon-Meunier, B. (2016). Interpretability of fuzzy linguistic summaries. *Fuzzy Sets and Systems*, 292, 307–317.

Schield, M. (2011). Statistical literacy: a new mission for data producers. *Statistical Journal of the IAOS*, 27, 173–183.

Skowron, A., Jankowski, A., Swiniarski, R.W. (2015). Foundations of rough sets. In: Kacprzyk, J., Pedrycz, W. (Eds.), *Handbook of Computational Intelligence*. Springer, Berlin, Heidelberg, pp. 331–348.

Smits, G., Pivert, O., Hadjali, A. (2014). Fuzzy cardinalities as a basis to cooperative answering. In: Pivert, O., Zadrożny, S. (Eds.), *Flexible Approaches in Data, Information and Knowledge Management. Studies in Computational Intelligence*, Vol. 497. Springer, Cham, pp. 261–289.

Terán, L., Kaskina, A., Meier, A. (2016). Maturity model for cognitive cities. In: Finger, M., Portmann, E. (Eds.), *Towards Cognitive Cities – Advances in Cognitive Computing and its Application to the Governance of Large Urban Systems*. Springer, Cham, pp. 35–59.

Vaisman, A., Zimányi, E. (2014). *Data Warehouse Systems – Design and Implementation*. Springer-Verlag, Berlin, Heidelberg.

Wang, T-C., Lee, L.H-D., Chen, C-M. (2007). Intelligent queries based on fuzzy set theory and SQL. In: *39th Joint Conference on Information Science*, Salt Lake City, pp. 1426–1432.

Yager, R.R. (1982). A new approach to the summarization of data. *Information Sciences*, 28, 69–86.

Yager, R.R. (1984). General multiple-objective decision functions and linguistically quantified statements. *International Journal of Man-Machine Studies*, 21, 389–400.

Yager, R.R. (1988). On ordered weighted averaging operators in multicritera decision making. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-18, 183–190.

Yager, R.R., Ford, M., Canas, A.J. (1990). An approach to the linguistic summarization of data. In: *3rd International Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems*, IPMU 1990, Paris, pp. 456–468.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, 8, 338–353.

Zadeh, L.A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, 9, 149–184.

P. Sojka is an assistant at the University of Economics in Bratislava, Faculty of Economic Informatics (Slovakia) since 2016. His work is mainly focused on the administration of operating systems and database systems, and also on programming applications for the teaching and research purposes. Before entering the University, he worked as a system/database administrator at the State Treasury and Debt and Liquidity Management Agency, then as an information security specialist in the Agricultural Paying Agency.

M. Hudec is an associate professor at the University of Economics in Bratislava, Faculty of Economic Informatics (Slovakia) and VSB – Technical University of Ostrava, Faculty of Economics (Czech Republic). He received the PhD degree from the University of Belgrade (Serbia). His work is mainly focused on fuzzy logic, knowledge discovery and information systems. He is a member of program committees of several international conferences and serves as an associate editor in two journals. He has published more than 50 articles including a monograph in Springer.

M. Švaňa is a PhD student at the VSB – Technical University of Ostrava, Faculty of Economics (Czech Republic). His work is focused mainly on the application of machine learning and natural language processing methods in decision making and economics, the topic of his dissertation being the relationship of sentiment on social networks and financial risk/uncertainty. His previous experiences in commercial sector include data analyst and web developer positions in several companies.