Merging Validity and Coverage for Measuring Quality of Data Summaries

Miroslav Hudec

Abstract Data summarization by quantified sentences of natural language simulates human reasoning in summing up from the data. Linguistic summaries are focused either on a whole data set, or on a part of a data set delimited by the flexible restrictions expressed as fuzzy sets. First, the paper examines influences of t-norms in compound predicates merged by the *and* connective and constructed fuzzy sets on the validity (truth value) of summaries. Further, linguistic summaries with restriction may express mined knowledge from the outliers and therefore be of low quality, even though the validity of summary could be high. The main aim of this paper is building a quality measure based on validity and coverage. Finally, additional possibilities related to the suggested measure and perspective topics for future research are outlined.

Keywords Linguistic summaries · Validity · Quality · Outliers · T-norms · Fuzzy sets

1 Introduction

Nowadays, mining summarized information from data sets is a topic of interest for researchers and practitioners. Data summarization can be efficiently realized by statistical methods which however, are understandable for rather small group of specialists. This observation is expressed in [1] as: "summarization would be especially practicable if it could provide us with summaries that are not as terse as the mean". Graphical interpretation is a valuable way of summarization but cannot be always effective [2]. Linguistics is an interesting alternative when data is hard to show graphically [3]. A linguistically summarized sentence can be read out by a text-to-speech synthesis system. It especially holds when the visual attention should

M. Hudec (✉)
Faculty of Economic Informatics, University of Economics in Bratislava, Bratislava, Slovakia
e-mail: miroslav.hudec@euba.sk

P. Kulczycki et al. (eds.), *Information Technology and Computational Physics*, Advances in Intelligent Systems and Computing 462,
DOI 10.1007/978-3-319-44260-0_5

not be disturbed [4]. These advantages hold when the resulting summarization is of a high quality.

People tend to summarize by terms of natural language. But, literally unlimited variations of linguistic terms and their modifications for expressing summaries exist. In order to put together mathematical formalization and people's preferred way, quantified sentences of natural language, i.e. Linguistic Summaries (LSs) were introduced in [5]. Since then LSs have been intensively researched in, e.g. [6–15].

Generally, LSs summarize the whole data set or a restricted part. In the former, LSs are of the structure *Q entities are (have) S*, where *Q* is a quantifier, and *S* is a summarizer. One example of such a summary is: most of houses have high gas consumption. In the latter, LSs are of the structure *Q R entities are (have) S*, where *R* puts some restriction on data sets. One example of such a summary is: most of old houses have high gas consumption. In addition, *R* and *S* can be consisted of several atomic predicates merged by the *and* connective [7, 8] which is usually modelled by t-norms [16]. The truth value of LSs (also called validity) gets value form the [0, 1] interval by agreement. Hence, validity is influenced by selected t-norm and constructed fuzzy sets.

LSs with restriction may be trapped into outliers due to possible very low coverage of tuples in *R* and *S* parts, even though the validity is high. Hence, this problem of the LSs quality should not be neglected. Hirota and Pedrycz [17] suggested five quality measures: validity, generality, usefulness, simplicity and novelty. These measures are further examined for LSs with the restriction part in [15] for the purpose of converting mined summaries into fuzzy rules. Further set of measures was introduced in [18, 19].

The main goal of this paper is focused on building outlier measure expressed by coverage and validity [15, 17]. Preliminary results in this direction were published in [20]. Furthermore, this paper extends discussion to the influence of t-norms and fuzzy sets to the validity of LSs. The reminder of this chapter is organized as follows. Section 2 gives some preliminaries of LSs which are used as a basis for the next sections. In Sect. 3 influences of different t-norms in *R* and *S* parts on validity are examined. Impact of constructed fuzzy sets is examined in Sect. 4. Section 5 is devoted to building a new quality measure related to outliers, discussion supported by illustrative example and future challenges. Section 6 gives a short note to different applications. Finally, Sect. 7 concludes this work.

2 Linguistic Summaries in Brief

LSs summarize knowledge from the data into the concise and easily understandable way for people. LS for summarizing the whole data set is of the structure *Q entities in database are (have) S*, where *Q* is a relative quantifier and *S* is a summarizer. Both are expressed by linguistic terms (fuzzy sets). The validity of summary is computed in the following way [5]:

$$v(Qx(Px)) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_S(x_i)\right) \tag{1}$$

where n is the number of tuples in a data set (cardinality), $\frac{1}{n}\sum_{i=1}^{n}\mu_S(x_i)$is the proportion of tuples in a data set that satisfy predicate S and μ_Q is the membership function of chosen relative quantifier.

LS with restriction has the form *Q R entities in database are (have) S*, where R is a restriction (expressed by fuzzy set) focusing on a part of data set relevant for the summarization task. The validity is computed in the following way [14]:

$$v(Qx(Px)) = \mu_Q\frac{\sum_{i=1}^{n} t(\mu_S(x_i), \mu_R(x_i))}{\sum_{i=1}^{n}(\mu_R(x_i)} \tag{2}$$

where $\frac{\sum_{i=1}^{n} t(\mu_S(x_i), \mu_R(x_i))}{\sum_{i=1}^{n}\mu_R(x_i)}$ is the proportion of tuples in a data set that satisfy S and belong to R, t is a t-norm and μ_Q is the membership function of chosen relative quantifier.

Linguistic terms such as *medium* (*around m*), *small* and *high* used in S and R can be expressed by triangular or trapezoidal fuzzy sets, L fuzzy set and linear gamma fuzzy set consequently (Fig. 1) ensuring the smooth transition between relevant and non-relevant tuples.

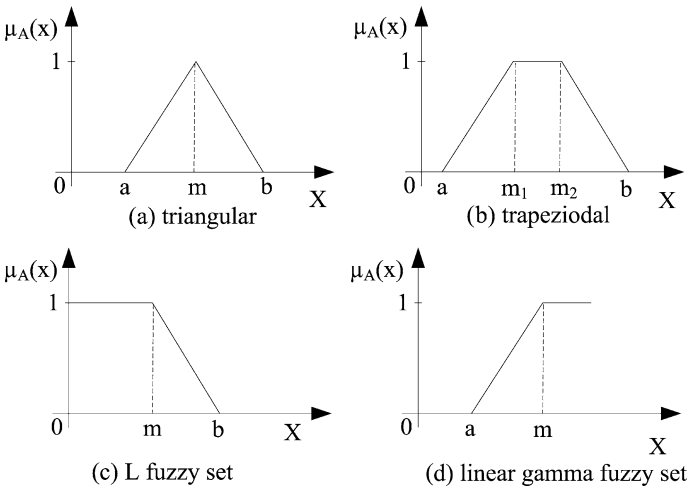

Fig. 1 Fuzzy sets for restrictions and summarizers

Summarizer and restriction may contain several atomic predicates merged by the *and* connective [7, 8]. These connectives are usually modelled by t-norms [16]. Four basic t-norms are:

- minimum t–norm:

$$t_m(\mu_{A_1}(x), \mu_{A_2}(x)) = \min(\mu_{A_1}(x),\ \mu_{A_2}(x)) \tag{3}$$

- product t-norm:

$$t_p(\mu_{A_1}(x), \mu_{A_2}(x)) = \mu_{A_1}(x) \cdot \mu_{A_2}(x) \tag{4}$$

- Łukasiewicz t-norm:

$$t_L(\mu_{A_1}(x), \mu_{A_2}(x)) = \max(\mu_{A_1}(x),\ \mu_{A_2}(x) - 1,\ 0) \tag{5}$$

- drastic product

$$t_d(\mu_{A_1}(x), \mu_{A_2}(x)) = \begin{cases} \min(\mu_{A_1}(x), \mu_{A_2}(x)) & \max(\mu_{A_1}(x), \mu_{A_2}(x)) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\mu_{Aj}(x)$ ($j = 1, 2$) denotes the membership degree to the j-th fuzzy set for element x.

The validity of LS is computed by the relative quantifiers such as *few, about, half, most of*. The *most of* quantifier, plotted in Fig. 2, is often used because users are interested to see which summaries are met by the majority of tuples.

We can say that the linguistic summary is

a more or less accurate textual description (summary) of a data set

This simple definition hides many challenges: construction of fuzzy sets for summarizers, restrictions and quantifiers, selecting appropriate t-norms, sufficient coverage of data, simplicity, usefulness, accuracy, summarizing from the outliers instead from the regular data and the like. The influences of t-norms, construction of fuzzy sets, coverage and outliers to quality of LSs are examined in the next sections.

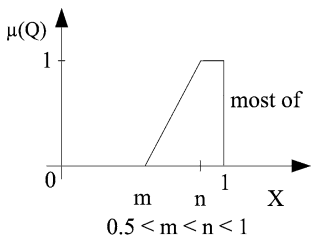

Fig. 2 Relative quantifier *most of*

3 Impact of T-Norms in Restriction and Summarizer to Validity

For LSs with restriction the quality is measured for each data point x_i ($i = 1, \ldots, n$) by t-norm in the numerator of (2) [2]. This section is focused on searching for suitable t-norms not only for merging restriction and summarizer but also for conjunction of atomic predicates inside restriction and summarizer. Two examples of such queries are: most high polluted and low situated (altitude) municipalities have a high number of respiratory diseases, and most middle aged customers have high turnover and small payment delays.

When restriction or summarizer consists of several atomic conditions (predicates P) connected by the *and* operator, t-norms come to the stage. All t-norms meet all axiomatic properties explained in e.g. [22], but differ in satisfying algebraic properties. Let us recall the following three algebraic properties [16]:

- The t-norm is an idempotent one if for $\forall a \in [0, 1],\ t(a, a) = a$
- The t-norm is a nilpotent one if there exists some $n \in N$ such that $t^{(n)}(a) = 0$
- The t-norm has a limit property if for $\forall a \in (0, 1),\ \lim_{n \to \infty} t^{(n)}(a) = 0$

LSs express proportion of tuples which meet atomic or compound predicate in S and/or R. For instance, when each atomic predicate P_j ($j = 1, \ldots, n$) is satisfied with degree of 0.48, then the tuple should participate in S with degree of 0.48. This requirement meets idempotent t-norm. The only idempotent t-norm is the minimum one (3). Furthermore, this t-norm is not nilpotent and does not have limit property. Łukasiewicz t-norm (5) meets the second property causing that tuple participates in proportion with value of 0. Product t-norm (4) meets third property causing decreasing tuples participation in the proportion, when the number of atomic predicates increases. When $j = 2$, tuple participates with degree of 0.2304; but when $j = 4$, tuple participates in summary with degree of 0.05308.

For the basic structure of LSs (1) when S is a compound predicate selecting the suitable t-norm is a pivotal task for obtaining LS of a high quality. Concerning the LS with restriction (2), selecting appropriate t-norm influences quality but further quality aspects should be considered.

To summarize, the only suitable t-norm is the minimum one (3), because it does not unnaturally reduce the proportion of tuples in a data set that satisfy LS. Interestingly, in the Sect. 5 the situation regarding suitable t-norms is opposite.

4 Influence of Constructed Fuzzy Sets to Validity and Coverage

The subjectivity in constructing fuzzy sets may influence quality of summarized information. It especially holds for the sufficient coverage and outliers which are examined later on.

The domains of attributes are, during the database design phase, defined in a way that all theoretically possible values can be stored. For instance, for the attribute monitoring frequency of an activity during a year the domain is the [0, 365] interval of integers. In practice, collected values can be far from the lower and upper limits of the domain. In the constructing fuzzy sets this fact should be considered [23], because users are not always aware of collected attributes' values. The situation plotted in Fig. 3a, where *L* and *H* are the lowest and the highest values in the current content of attributes, respectively, and *Dmin* and *Dmax* are the lower and upper limit of domains, respectively, might appear. The truth value equal to 1 in Fig. 3a may express summary on outliers and therefore, is of a low quality.

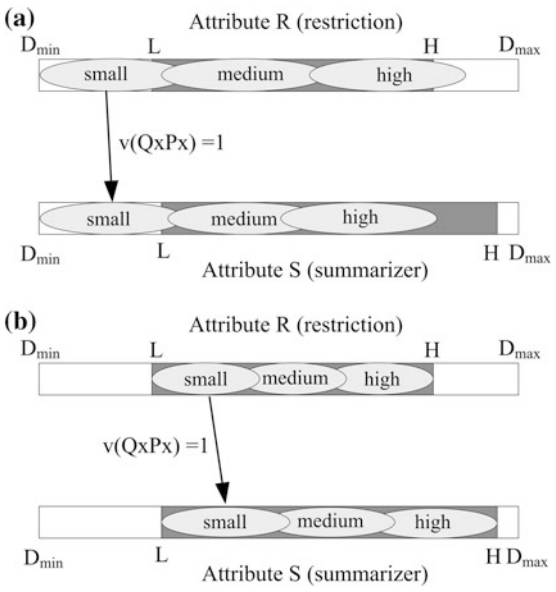

Fig. 3 Fuzzy sets for restriction and summarizer: **a** fuzzy sets do not reflect stored data; **b** fuzzy sets reflect stored data

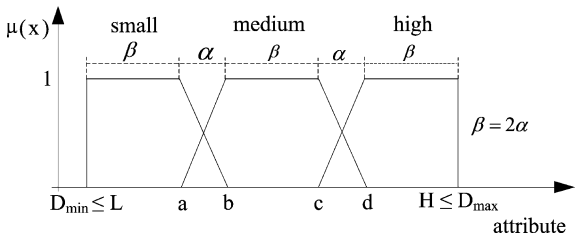

Fig. 4 Fuzzy sets uniformly distributed in the part of attribute domain covered by data

Moreover, one should be very careful when no tuple meets the *R* part because it leads to dividing by zero in (2).

In order to mitigate this problem, we should construct membership functions considering only parts of domains that contain data [23]. The validity equal to 1 in Fig. 3b can be relevant summary. But it does not hold automatically.

The shapes of membership functions have adopted several conventions [24]. Generally, the membership functions are convex and normalized piecewise linear functions. Figure 4 shows the situation where the family of fuzzy sets consists of three sets to cover terms *small*, *medium* and *high*. The flat segments of these fuzzy sets (β) express no uncertainty in belonging to sets, whereas parameter α expresses the uncertainty in belonging to a set. When $\alpha = 0$, the domain is partitioned into crisp sets. If a requirement for finer granulation exists, more fuzzy sets (e.g. five sets: very small, small, medium, high, very high) can be straightforwardly constructed adjusting parameters α and β. These concepts can be defined by nonlinear functions as well. Concerning practical applications and the simplicity for end users, linear functions are often preferable.

Even though fuzzy sets are constructed on parts of domains where data are recorded, the data distribution far from the uniform one might cause that LSs express relations detected in outliers. For example, let only 20 of $5 \cdot 10^6$ tuples fully meet the *R* and the same tuples fully meet the *S*, then the validity (2) gets the value of 1, leading us to the false conclusion.

5 Quality Measure Focused on Outliers and Coverage

Keeping the aforementioned in mind, we can say that if LSs with restriction have high validity *v* (2), it does not straightforwardly mean that these LSs are suitable for expressing summarized information, even though suitable t-norm is applied and care was taken during the construction of fuzzy sets. Thus, quality measures should be applied in order to mitigate vagueness of calculated validity. Five quality measures: validity, generality, usefulness, novelty and simplicity were suggested in [17] and further examined in [15]. Four measures: coverage, brevity (or shortness), specificity and accuracy mainly for non-quantified linguistic summaries are examined in [18]. All these measures get values from the [0, 1] interval.

The novelty measure means that unexpected summaries represent valuable knowledge, if they do not express knowledge mined from the outliers [15] (errors in observations or existence of few very different tuples). Therefore, for calculating the novelty measure outliers should be recognized and measured. Furthermore, outliers and coverage are related. The outlier's measure is examined in this section.

5.1 Outliers

Wu et al. [15] explained that outliers appear if the validity degree v is very small or very high and the sufficient coverage C must be very small. Therefore this measure can be expressed as

$$O = \min(\max(v, 1-v), (1-C)) \tag{7}$$

where C is the coverage, which is defined later. If coverage is small ($C \rightarrow 0$), then outlier measure O is near the value of 1 (if v gets value near 1 or 0). If coverage is high ($C \rightarrow 1$), then the outlier measure is near the value of 0. In a general way (7) can be expressed as:

$$O = t(s(v, 1-v), (1-C)) \tag{8}$$

where t is a t-norm and s is a s-norm.

The non-outlier measure is calculated as the negation of (8) by De Morgan's law, i.e.:

$$1 - O = s(t(1-v, v), C) \tag{9}$$

when the standard fuzzy negation is used.

We can say that LSs are of a high quality if validity and non-outliers are high. This observation is formally written as

$$Q_c = t(v, 1-O) = t(v, s(t(1-v, v), C)) \tag{10}$$

From the properties of t-norms holds: $t(1 - v, v) \leq 0.5$. If we define quality as significant, when coverage is higher or equal 0.5, then from (10) yields:

$$Q_c = \begin{cases} t(v, C) & C \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $C = 0.5$ is considered as a threshold value of coverage.

The next task is calculating coverage in a way that it meets the requirement (11).

Coverage

Concerning the basic structure of LS (1), the whole data set is covered due to appearance of the variable n (cardinality of a data set) in the denominator, i.e. coverage is implicitly calculated. If the coverage is low, then it directly influences the validity. Regarding the LS with restriction (2), coverage should be calculated explicitly. The following coverage index for LSs of structure (2) is created [25]:

$$i_C = \frac{\sum_{i=1}^{n} t(\mu_S(x_i), \mu_R(x_i))}{n} \tag{12}$$

where n is the number of tuples in a data set. Other variables have the same meaning as in (2). The coverage index i_c explains how many records' membership degrees influence the validity of a LS. In practice, the coverage index is a small number, because LSs with restriction usually cover relatively small subset of the considered data set [15]. Therefore, the mapping which converts i_c (12) into the coverage C (used in (7–11)) yields [15]:

$$C = f(i_C) = \begin{cases} 0, & i_c \le r_1 \\ 2\left(\frac{i_c - r_1}{r_2 - r_1}\right)^2, & r_1 \le i_c < \frac{r_1 + r_2}{2} \\ 1 - 2\left(\frac{r_2 - i_c}{r_2 - r_1}\right)^2, & \frac{r_1 + r_2}{2} \le i_c < r_2 \\ 1, & i_c \ge r_2 \end{cases} \tag{13}$$

where $r_1 = 0.02$ and $r_2 = 0.15$. Anyway, parameters r_1 and r_2 can be set according to user preferences in a same way as for other fuzzy sets: for S and R (Figs. 1 and 4) and quantifiers (Fig. 2). When R is more restrictive, i.e. several atomic predicates merged by the *and* connective, then parameters r_1 and r_2 can be smaller.

Naturally, the question which t-norm in (11) is the suitable one appears. Let us have calculated values of validity and coverage for two LSs shown in Table 1.

The minimum t-norm (3) says that *ls1* and *ls2* are indistinguishable. Hence, we need t-norm which considers all attributes, not only attributes bearing minimal value. The solution provides product t-norm (4) stating that *ls1* is of higher quality than *ls2* (Table 1). It is the opposite observation than for aggregating atomic predicates by the *and* connective in S (1), (2) and R (2) parts of LSs (Sect. 3). Instead of product t-norm, we can apply another non-idempotent t-norm: a Łukasiewicz one, but it further decreases measure (11).

Table 1 Merging validity and coverage of LSs by product and minimum t-norms in (11)

LS	Validity	Coverage	$Q_c(v, C)$ by (4)	$Q_c(v, C)$ by (3)
ls1	0.75	0.95	0.7125	0.75
ls2	0.75	0.75	0.5625	0.75

The quality measure regarding the outliers can be also expressed as

$$Q_c = f(v, 1 - C) \tag{14}$$

In this case, we do not consider coverage (13) but its negation. This equation represents a bipolar relation because v is a positive predicate and $(1 - C)$ is a negative one.

Furthermore, if the requirement for a high quality is the full coverage (13), i.e. $C = 1$, then the non-continuous drastic t-norm (6) is a rational option for merging validity and coverage:

$$Q_c = d_p(v, C) = \begin{cases} v, & C = 1 \\ C, & v = 1 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

The validity of the LS is taken into account only if $C = 1$. All summaries which pass this filter can be ranked downwards form the best one according to the validity degree. Summaries are evaluated by their respective validities, as is the case in (2), but only when they pass this simple filter.

Illustrative Example

In order to mine all relevant summaries, user has defined set of attributes, quantifiers and linguistic terms for attributes appearing in restriction and summarizer. For simplicity, mined LSs are written as *ls1* (i=1,…,9) and shown in Table 2.

When coverage is fully satisfied the same result is obtained by (11) and (15). The latter is a filter and expressed as first meet coverage and then validity. This approach is suitable when coverage is a sharp condition. Otherwise, product t-norm is the option. The drawback of drastic product is in its sharpness. When both validity and coverage are close to 1, the result is 0.

However, in measuring quality by drastic product (15) we have the second option: when $v = 1$, the result is C (the last record in Table 2). This option may be either excluded or used as an alternative: if validity is fully satisfied, then preferable summary is one with higher coverage degree.

Table 2 Quality of mined LSs

LS	Validity	Coverage	$Q_c(v, C)$ (11) by product t-norm	Q_c (15)
ls1	0.80	0.80	0.6400	0.00
ls2	0.75	0.85	0.6375	0.00
ls3	0.65	1.00	0.6500	0.65
ls4	0.93	1.00	0.9300	0.93
ls5	0.81	0.24	0.0000	0.00
ls6	0.12	1.00	0.1200	0.12
ls7	0.23	0.14	0.0000	0.00
ls8	0.95	0.95	0.9025	0.00
ls9	1.00	0.58	0.5800	0.58

Mining LSs from the Data

Generally, two ways for mining summaries exist:

- User defines all relevant linguistic terms for quantifiers, restrictions and summarizers and all attributes of interest.
- User defines term sets for quantifiers, summarizers and restrictions without selecting relevant attributes.

In both cases an application reveals all summaries for which validity (1) or (2) is higher than 0, or higher than the defined threshold value. The difference is in mined summaries. In the first way, only summaries of a clear interest are mined. In the next step quality measure (11) can be applied. In the second way, the usefulness of mined summaries is a further measure which should be considered, i.e. high validity and coverage of a quantified sentence: most territorial units with high percentage of public greenery have small unemployment, presumably is irrelevant for analysing reasons for high unemployment and building related rule base.

Some Perspectives for Further Research

The first perspective is aggregating quality measures mentioned in [15, 18] and measure suggested in this work. But it is not an easy task because we need to aggregate several measures which may be partially redundant and conflicting [25].

For instance, the simplicity measure [15] concerns the syntactic and semantic complexity of the LSs. This measure expresses how many attributes in restriction and summarizer in a summary exist. Complex summaries are less legible for users. Hence, the simplicity measure can be expressed as [15]:

$$S_{im} = 2^{2-l} \tag{16}$$

where l is a total number of atomic predicates in restriction and summarizer. Evidently, Sim gets values from the unit interval. The example of a summary having $Sim = 1$ is: most young customers have a small payment delay.

Regarding the basic structure of LS (1), Eq. (16) yields:

$$S_{im} = 2^{1-l} \tag{17}$$

ensuring that the simplest structure (one atomic predicate inside the summarizer, e.g. most customers are middle aged) has simplicity equal to 1.

The second perspective is the focus on quantified restrictions and summarizers. A structure of LSs with restriction (2) can be also expressed as

$$Q\left(\wedge_{i=1}^{n} R_i(x) \right) \text{are} \left(\wedge_{j=1}^{m} S_j(x) \right) \tag{18}$$

where R_i and S_j are atomic predicates in restriction and summarizer consequently.

When $i = j = 1$ we obtained the structure frequently examined in the literature. It is obvious that when n and m are larger numbers the sentence becomes very restrictive. The structure (18) can be relaxed to the following structure:

$$Q\left(\text{most of } R_i(x), i = 1, \ldots, n\right)\text{are}\left(\text{most of } S_j(x), j = 1, \ldots, m\right) \tag{19}$$

This structure corresponds with the structure of quantified queries [26], where tuples which meet the majority of atomic predicates are selected.

The benefit is a less restrictive summary concerning all atomic predicates. A tuple which meets four atomic predicates with degrees 0.2, 0.1, 0.25, 0.2 has a lower impact than a tuple which meets these predicates with degrees 1, 0.95, 0.9, 0. Drawback lies in the fully non-satisfied predicate. Attribute's value might be very far from the acceptable value or very close. Apparently, this is a challenge for future research where the cardinalities of tuples which are in predicates' neighbourhoods should be measured. The calculation of validity is not as complex task as coverage, because validity is directly calculated from (19).

6 Short Note to Applications

LSs are applicable in a variety of tasks. Three of them are mentioned in this section. Presumably, the first attempt to apply LSs with restriction in data imputation related to the item non-response was discussed in [27]. For this purpose we need to calculate validity (2) by the more restrictive quantifier *most of*. The restriction is realized by adjusting parameters of the quantifier shown in Fig. 2 in the following way: $m > 0.5$ and $n = 1$ yielding the quantifier *almost all*. Further, when validity is significant but not sufficiently high we should focus on a more restrictive part of a database. One option is the conjunction of initial and additional atomic predicates in the R part. Hence, the care should be taken when constructing fuzzy sets. Further, a minimum t-norm should be used for merging atomic predicates. Finally, quality measures should be applied. Regarding quality measures, validity and coverage are more important than simplicity. A more restrictive part of a database may have strong relation between attributes (high values of validity and coverage) but the simplicity measure (16) is low. Therefore, in the terms of bipolar approaches validity and coverage (11) are restrictions and simplicity is desire. In this way LSs might be competitive to other data imputation approaches but definitely further research is required.

The second method of application is converting mined LSs into fuzzy if-then rules [15, 23]. The research of quality measures was influenced by this task, because fuzzy rules should be of a high quality due to their broad applicability in, e.g. control and classification. Therefore, the aforementioned aspects of an LSs quality should not be neglected. Furthermore, less complex rules are preferred. Hence, the simplicity measure (16) has in this field higher importance than in the data imputation field.

The third kind of applications is mining "abstracts" from the data for informative purposes and to support decision and policy making processes. In the former, the less restrictive quantifier *majority of* can be applied. It corresponds with the *most of* quantifier defined in [9] as $m = 0.3$ and $n = 0.85$ (Fig. 2). Other quality aspects should be also considered depending of the type of LS. Contrary to the two aforementioned types of tasks, in these tasks both types of LSs (basic structure and structure with restriction) are applicable. In this field reading LSs by a text-to-speech synthesis system is a suitable way for distributing mined information to users. Thus, the simplicity is a measure which should have similar importance as validity and coverage.

Although these three kinds of tasks are used for different purposes (from data collection through data analysis to data dissemination), they share quality issues but different relevance of particular quality measures.

7 Concluding Remarks

LSs play a pivotal role in summarizing information from the data when uncertainty related to the semantic meaning of the phenomena (fuzziness) is included in the task. The validity of the LS may be influenced by constructed fuzzy sets, or selected t-norm function, or may explain relational knowledge in outliers. The last observation holds for LSs with restriction part (2). Outliers appear due to the measurement and observational errors and when very few tuples has significantly different values than the high majority of tuples. At any rate, before accepting LSs, it is advisable to filter them by quality measure(s).

In this chapter, we have created a simplified outlier measure that consists of coverage and validity merged by t-norm. LS is of a sufficient quality if it has high validity and high coverage. The suggested quality measure (11) can be used as a standalone one when non-outlier coverage and validity are sufficient. Furthermore, this measure can be part of the set of quality measures. As a connective in this measure the minimum t-norm should be avoided. Suitable t-norms are those which take into consideration both attributes and do not meet idempotency property. Hence, the option is product t-norm. In cases when the full coverage ($C = 1$) is required, the suitable connective can be obtained by drastic t-norm. In this case all summaries which pass this simple filter can be ranked according to the validity degree.

Concerning the *and* connective in compound restriction and summarizer, we believe that the only suitable t-norm is the minimum t-norm, because the proportion of tuples which contribute to the summary is not unnaturally decreased.

In the future activities, we will focus on aggregating quality measures into the compound one and on developing quality measures for summaries consisted of quantified restriction and summarizer.

References

1. Yager, R.R., Ford, M., Cañas, A.J.: An approach to the linguistic summarization of data. In: 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU '90), Paris, France, July 2–6, pp. 456–468 (1990)
2. Lesot, M.-J., Moyse G., Bouchon-Meunier, B.: Interpretability of fuzzy linguistic summaries. Fuzzy Sets Syst. (In press) **292**(1), 307–317 (2016)
3. Yu, J., Reiter, E., Hunter, J., Sripada, S.: Sumtime-turbine: a knowledge-based system to communicate gas turbine time-series data. In: Chung, P.W.H., Hinde, C.J., Ali, M. (eds.) Lecture Notes in Computer Science, LNAI, vol. 2718, pp. 379–384. Springer, Berlin, Heidelberg (2003)
4. Arguelles, L., Triviño, G.: I-struve: automatic linguistic descriptions of visual double stars. Eng. Appl. Artif. Intell. **26**(9), 2083–2092 (2013)
5. Yager, R.R.: A new approach to the summarization of data. Inf. Sci. **28**(1), 69–86 (1982)
6. Bouchon-Meunier, B., Moyse, G.: Fuzzy linguistic summaries: where are we, where can we go? In: 2012 IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFEr 2012), New York, USA, March 29–30, pp. 1–8 (2012)
7. 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. PhysicaVerlag, Heidelberg (1996)
8. Hudec, M.: Issues in construction of linguistic summaries. In: Mesiar, R., Bacigál, T. (eds.) Proceedings of Uncertainty Modelling 2013, pp. 35–44. STU, Bratislava (2013)
9. Kacprzyk, J., Zadrożny, S.: Protoforms of linguistic database summaries as a human consistent tool for using natural language in data mining. Int. J. Software Sci. Comput. Intell. **1**(1), 1–11 (2009)
10. Kacprzyk, J., Yager, R.R.: Linguistic summaries of data using fuzzy logic. Int. J. General Syst. **30**(2), 133–154 (2001)
11. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of time series using a fuzzy quantifier driven aggregation. Fuzzy Sets Syst. **159**(12), 1485–1499 (2008)
12. Niewiadomski, A., Ochelska, J., Szczepaniak, P.S.: Interval-valued linguistic summaries of databases. Control Cybern. **35**, 415–443 (2006)
13. Raschia, G., Mouaddib, N.: SAINTETIQ: a fuzzy set-based approach to database summarization. Fuzzy Sets Syst. **129**(2), 137–162 (2002)
14. Rasmussen, D., Yager, R.R.: Summary SQL—A fuzzy tool for data mining. Intell. Data Anal. **1**(1–4), 49–58 (1997)
15. Wu, D., Mendel, J.M., Joo, J.: Linguistic summarization using if-then rules. In: 2010 IEEE International Conference on Fuzzy Systems, Barcelona, Spain, July 18–23, pp. 1–8 (2010)
16. Klement, E.P., Mesiar, R., Pap, E.: Triangular norms: basic notions and properties. In: Klement, E.P., Mesiar, R. (eds.) Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms, pp. 17–60. Elsevier, Amsterdam (2005)
17. Hirota, K., Pedrycz, W.: Fuzzy computing for data mining. Proc. IEEE **87**(9), 1575–1600 (1999)
18. Castillo-Ortega, R., Marín, N., Sánchez, D., Tettamanzi, A.: Quality assessment in linguistic summaries of data. In: 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2012), Catania, Italy, July 9–13, pp. 285–294 (2012)
19. Pereira-Fariña, M., Eciolaza, L., Triviño, G.: Quality assessment of linguistic description of data. In: ESTYLF, Valladolid, Spain, February 1–3, pp. 608–612 (2012)
20. Hudec, M.: Merging validity and coverage for measuring quality of data summaries. In: Congress on Information Technology, Computational and Experimental Physics, Cracow, Poland, December 18–20, pp. 149–153 (2015)
21. Zadrożny, S., Kacprzyk, J.: Issues in the practical use of the OWA operators in fuzzy querying. J. Intell. Inf. Syst. **33**(3), 307–325 (2009)

22. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Applications. Academic Press, New York (1980)
23. Hudec, M., Vučetić, M., Vujošević, M.: Synergy of linguistic summaries and fuzzy functional dependencies for mining knowledge in the data. In: 18th International Conference on System Theory, Control and Computing (IEEE ICSTCC), Sinaia, Romaina, October 17–19, pp. 335–340 (2014)
24. Garibaldi, J.M., John, R.I.: Choosing membership functions of linguistic terms. In: 12th IEEE International Conference on Fuzzy Systems (FUZZ '03), St. Louis, USA, May 25–28, pp. 578–583 (2003)
25. Hudec, M.: Linguistically summarizing hierarchical data. In: 16th IEEE International Symposium on Computational Intelligence and Informatics (CINTI 2015), Budapest, Hungary, November 19–21, pp. 141–145 (2015)
26. Kacprzyk, J., Ziółkowski, A.: Database queries with fuzzy linguistic quantifiers. IEEE Trans. Syst. Man Cyber. SMC-**16**(3):pp. 474–479 (1986)
27. Hudec, M.: Linguistic summaries applied on statistics—case of municipal statistics. Austrian J. Stat. **43**(1), 63–75 (2014)