Proceedings of the 2013 Federated Conference on
Computer Science and Information Systems pp. 643–646

On some quality criteria of bipolar linguistic summaries

Mateusz Dziedzic
Department of Automatic Control
and Information Technology
Cracow University of Technology
ul. Warszawska 24, 31–155 Kraków, Poland
also
PhD Studies, Systems Research Institute
Polish Academy of Sciences
Email: Mateusz.Dziedzic@ibspan.waw.pl

Janusz Kacprzyk, IEEE Fellow
and Sławomir Zadrożny
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6, 01–447 Warszawa, Poland
Email: {Slawomir.Zadrozny,Janusz.Kacprzyk}@ibspan.waw.pl

***Abstract*—The quality measures for bipolar linguistic summaries of data, as proposed in our previous work [1], are further developed. The summaries introduced in [2] are assumed to be an extension of the "classical" linguistic summarization (cf. [3], [4]), a human-consistent data mining technique revealing complex patterns present in data. This extension consists in using the "and possibly" to build a summary and introducing the notion of context to determine the validity of the summary. We present a more detailed description of summaries quality measures/criteria and reports results of more extensive computational experiments.**

I. Introduction

THE AIM of data mining is to discover patterns in data in a form interesting and clear to the end user. A promising way to achieve this is to use (quasi) natural language. This has been a motivation for the *linguistic data summaries* introduced by Yager [3] and further developed by him [5] and other contributors, notably Kacprzyk and Zadrożny [6], [7].

Recently, an important role of bipolarity of user preferences, in particular in fuzzy linguistic querying [8], is noticed. Its essence is in considering both positive and negative evaluations of objects in question which are not necessarily complements of each other. An important and most interesting line of research focuses on the treatment of negative evaluations as obligatory while the positive evaluations as somehow secondary. This results in the introduction and study of the "and possibly" logical connective [9]. Moreover, the concept of bipolar queries involving such a connective has been proposed [10] to better model user preferences as exemplified by the query "Find a house, cheap *and possibly* located close to a station".

In our previous papers [1], [2] we began to study if relation between fuzzy linguistic queries and linguistic data summaries may be adopted for bipolar queries. The results were positive and led us to the concept of bipolar linguistic summaries of data. In this paper we focus on two quality criteria of such new type of linguistic summaries, introduced in [1] and referring to the notion of the context of a summary.

The structure of the paper is as follows. In Section II we briefly remind the basics of the fuzzy linguistic queries and "classical" linguistic summaries, and introduce the notation to be used in the rest of the paper. In Section III we discuss the concepts of bipolar queries and bipolar linguistic summaries. Section IV reports on the computational experiments focused on comparing different summary contexts and discusses the results obtained.

II. Fuzzy linguistic queries and linguistic data summaries

A. Fuzzy linguistic queries

In classical query languages, such as SQL, preferences of users must be expressed precisely. However, due to the fact that their original form is a natural language expression, they are very often imprecise. For example, one may be concerned primarily with the cost while looking for an apartment to rent and express his or her preference as:

Find *cheap* apartments for rent in Kraków. (1)

In an approach, referred here to as fuzzy linguistic queries, such imprecise terms (e.g. *cheap*) are represented by fuzzy sets defined in the domains of respective attributes.

Usually, a dictionary of linguistic terms is assumed as a part of an implementation which contains predefined linguistic terms and corresponding fuzzy sets as well as terms defined by the users. Linguistic terms collected in a dictionary are a starting point to derive meaningful *linguistic summaries* of a database.

B. Linguistic summaries of data

As linguistic summaries we understand a (quasi) natural language sentences that grasp some characteristic features of data collected in a database. We use Zadeh's calculus of linguistically quantified propositions as the underlying formalism. The statement representing a linguistic summary points out some properties shared by a number of data items and the proportion of these data items is expressed using a *linguistic quantifier*. Yager [3], [5] first proposed the use of linguistically quantified propositions to summarize data in a user consistent

978-1-4673-4471-5/$25.00

way. That idea has been further developed, cf., e.g., Kacprzyk and Yager [11], and Kacprzyk, Yager and Zadrożny [4], [6].

Assuming $R = \{t_1, \ldots, t_n\}$ is a set of tuples (a relation) in a database, representing, e.g., a set of employees; $A = \{A_1, \ldots, A_m\}$ is a set of attributes defining schema of the relation R, e.g., salary, age, education_level, etc. in a database of employees ($A_j(t_i)$ denotes a value of attribute A_j for a tuple t_i), the linguistic summary of a set R is a linguistically quantified proposition which is an instantiation of one of the following abstract *protoforms* [12] of type I and type II, respectively:

$$Q_{t \in R}\ S(t) \tag{2}$$

$$Q_{t \in R}\ (U(t), S(t)) \tag{3}$$

then a linguistic summary is composed of the following elements: a *summarizer* S which is a fuzzy predicate representing, e.g., an expression "an employee is well-educated", formed using attributes of the set A; a *qualifier* U (optional) which is another fuzzy predicate representing, e.g., a set of "young employees"; a *linguistic quantifier* Q, e.g., "most" expressing the proportion of tuples satisfying the summarizer (optionally, among those satisfying a qualifier); *truth (validity)* T of the summary, i.e. a number from $[0, 1]$ expressing the truth of a respective linguistically quantified proposition.

In Yager's original approach [3] the linguistic quantifiers are represented using Zadeh's definition [13]. A *proportional*, *non-decreasing* linguistic quantifier Q is represented by a fuzzy set in $[0, 1]$ and $\mu_Q(x)$ states the degree to which the proportion of $100{\times}x$ % of elements of the universe match the proportion expressed by the quantifier Q. Thus, the truth degree of the linguistic summaries of type I (here we use only type I summaries, thus type II is omitted) is:

$$T(Q_{t \in R} S(t)) = Z_Q(S) = \mu_Q[\frac{1}{n} \sum_{i=1}^{n} \mu_S(t_i)] \tag{4}$$

III. Bipolar queries and bipolar linguistic summaries of data

A. Bipolar queries

In classical approaches to preferences modelling, notably in database querying, it is usually assumed that an alternative (tuple) is either accepted or rejected. However, the results of many studies, cf. [10], seem to suggest that the decision maker often comes up with somehow independent evaluations of positive and negative features of alternatives in question. This leads to a general concept of *bipolar query* against database, evaluation of which results in two degrees corresponding to the satisfaction of the positive and negative condition.

Most of the research on bipolar queries are focused on a special case where the positive and negative conditions are interpreted in an asymmetric way [10]. Namely, the latter is treated as a *constraint*, denoted C, which has to be satisfied, while the former plays the role of a mere *preference*, denoted P.

We follow the approach of Lacroix and Lavency [14], Yager [15], [16] and Bordogna and Pasi [9], adapted for database querying by Zadrożny and Kacprzyk [17], which combine both conditions using the "and possibly" operator which aggregates their satisfaction degrees depending on the possibility of a simultaneous matching of both conditions.

Thus, the bipolar query's condition may be formally written as:

$$C \text{ and possibly } P, \tag{5}$$

and may be illustrated with query: Find employees that are *young* and possibly earn a *high* salary. Such a bipolar query would be denoted (C, P) and interpreted as follows. If there is a tuple which satisfies both conditions, then and only then it is actually *possible* to satisfy both of them and each tuple of data has to do so, and, on the other hand, if there is no such a tuple, then condition P can be ignored. The matching degree of the (C, P) query against a tuple t may be formalized as [14]:

$$T\left(C(t)\ and\ possibly\ P(t)\right) = C(t) \wedge \left(\exists s\left(C(s) \wedge P(s)\right) \Rightarrow P(t)\right) \tag{6}$$

B. Bipolar linguistic summaries

The main idea behind the bipolar linguistic summaries is to relate the "and possibly" to *a part* of the database instead of the whole database. Let us consider the following example:

> *Most* employees have a *short* seniority *and, if possible with respect to similarly educated colleagues*, earn a *high* salary.

An employee matches such a summary if:

1) he or she has a short seniority (to a high degree) and earns a high salary (to a high degree), or
2) he or she has a short seniority (to a high degree) and there is no other *similarly educated* employee who earns a high salary.

A characteristic feature of such a summary is the use of a summarizer employing an extended version of the "and possibly" operator, which we will refer to as the "contextual and possibly" operator. This operator may be expressed as:

$$C \text{ and possibly } P \text{ with respect to } W. \tag{7}$$

For the purposes of bipolar queries (and, thus, bipolar linguistic summaries) the predicates C and P should be interpreted as the required and desired conditions, respectively, while the predicate W denotes the *context* in which the possibility of satisfying both C and P will be assessed, separately for each tuple. Then, the formula (7) is interpreted as:

$$T\left(C(t)\ and\ possibly\ P(t)\ with\ respect\ to\ W\right) = C(t) \wedge \left(\exists s\left(W(t,s) \wedge C(s) \wedge P(s)\right) \Rightarrow P(t)\right) \tag{8}$$

Our preliminary computational experiments show that usage of the standard De Morgan triples[1], both with the S- and

[1]As standard De Morgan triples we understand $(\wedge_{\min}, \vee_{\max}, \neg)$, $(\wedge_{\Pi}, \vee_{\Pi}, \neg)$ and $(\wedge_{\mathrm{L}}, \vee_{\mathrm{L}}, \neg)$ with t- and s-norms: Minimum $(\min(x, y))$ and Maximum $(\max(x, y))$; Product $(x \cdot y)$ and Probabilistic sum $(x + y - x \cdot y)$; and Łukasiewicz's $(\max(0, x + y - 1))$ and $(\min(1, x + y))$, respectively.

R-implication, in (8) may lead to somehow counter-intuitive results in terms of bipolar queries evaluation.

Thus we use the MinMax triple and Goguen $R-$implication which turns (8) into:

$$T\left(C(t)\ and\ possibly\ P(t)\ with\ respect\ to\ W\right) = \begin{cases} \min(C(t),1) & \text{for } \exists WCP(t)=0 \\ \min\left(C(t), \min\left(1, \frac{P(t)}{\exists WCP(t)}\right)\right) & \text{otherwise} \end{cases}, \tag{9}$$

where $\exists WCP(t)$ denotes $\max_{s\in R}\min(W(t,s),C(s),P(s))$.

C. Summary context quality criteria

In [1] we stated that the quality of the summary context $W(t,s)$ itself and the whole implication premise in (8) have to be considered when measuring the quality of the bipolar linguistic summaries.

Namely, if P and/or W are such that the premise of the implication in (8) is true to a very low or a very high degree *for most* of t's, then the summarizer (7) does not make much sense even if the truth value of a summary is high. This is due to the behaviour of the bipolar query "C and possibly P" which turns into "C" and "C and P", respectively, when the truth degree of $\exists_{s\in R}C(s)\wedge P(s)$ is close to 0 and close to 1.

The introduction of the context W partially alleviates this problem but W has to be chosen carefully. If for most t's there does not exist s such that $W(t,s)$, then the premise of the implication is most often false and the summary is true to a high degree for any P. We propose a solution to this problem in a form of quality measures expressed using the following linguistically quantified propositions:

$$Q_{t\in R}\exists_{s\in R\setminus\{t\}}W(t,s) \tag{10}$$

$$Q_{t\in R}\exists_{s\in R\setminus\{t\}}C(s)\wedge P(s)\wedge W(t,s). \tag{11}$$

Namely, if the truth of (10) for a summary is too small (lower than some threshold value), then such a summary should be discarded. Also, if the truth of (11) is too high (too close to 1; larger than the second threshold value) or too small (too close to 0; lower than the third threshold value), then the summary also shouldn't be taken into account. Obviously, if the first threshold is violated, then also the third one is. On the other hand, even if the first threshold is satisfied, the summary may still fail to satisfy thresholds two or three and should be discarded.

Tuple t is excluded from the range of the existential quantifiers in (10)–(11) as if the only tuple related via W with t is only t itself, then, naturally, the resulting summary is of no interest.

IV. Computational experiments and discussion

Data on the rates of return (RORs) of selected investment funds[2] (IFs) (Tab. I), are used to present examples of bipolar linguistic summaries and their semantics in scope of summary context quality.

[2]URL: http://www.analizy.pl/fundusze/ as of May 24, 2013.

Table I
Selected investment funds (IF)

No.	IF rating[a]	1-month ROR[b]	12-month ROR
1	5	-3.7	6.9
2	3	-0.8	20.8
3	2	-3.5	2.2
4	5	-3.5	2.5
5	5	-2.3	9.7
6	4	-2.2	9.5

[a] Rating from http://www.analizy.pl/fundusze/, as of May 24, 2013.
[b] Rate Of Return.

Table II
Investment funds (IF) - local neighbourhoods coefficients of tuples

	$W_1(t,s)$						$W_2(t,s)$						$W_3(t,s)$					
No.	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
1	1.0			1.0	1.0		1.0			1.0	1.0	1.0	1.0	1.0		1.0	1.0	1.0
2		1.0						1.0	1.0			1.0	1.0	1.0	1.0	1.0	1.0	1.0
3			1.0					1.0	1.0					1.0	1.0			1.0
4	1.0			1.0	1.0		1.0			1.0	1.0	1.0	1.0	1.0		1.0	1.0	1.0
5	1.0			1.0	1.0		1.0			1.0	1.0	1.0	1.0	1.0		1.0	1.0	1.0
6						1.0	1.0	1.0		1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Empty cell indicates no *similarity* ($W(t,s)=0.0$).

Fuzzy predicates are represented by trapezoidal membership functions and instantiated as (see Fig. 1):

- C: has "high"/"average"/"low" 12-month ROR,
- P: has "high"/"average"/"low" 1-month ROR,
- $W_{1/2/3}$: of "the same"/"very similar"/"quite similar" Rating, the former true iff IF rating(t) = IF rating(s) and the latter two defined over |IF rating(t) − IF rating(s)|.

In Tab. III we present summaries with truth value (evaluated using Zadeh's approach) $T>0$ obtained for data in Tab. I and compare the results in scope of proposed quality criteria (10) and (11).

In order to focus the interpretation on summaries contexts we consider only one linguistic quantifier Q with the membership function indicating the proportion of tuples satisfying the summarizer below 30% and above 80% as, respectively, totally

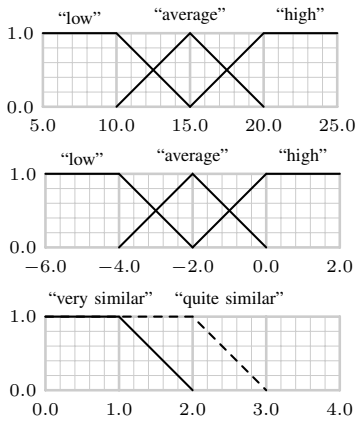

Figure 1. Membership functions of 12-month ROR (condition C – upper plot), 1-month ROR (condition P – center plot) and "similar" Rating (condition $W_{\text{Rating II/III}}$ – lower plot) predicates.

Table III
OBTAINED BIPOLAR LINGUISTIC SUMMARIES (USING ZADEH'S ($Z_{\mathbf{Q}}$) APPROACH)

No.	Linguistic summary (Q, C, P)	W_1 [a]	W_2 [a]	W_3 [a]
1	*"Most" of IFs have "low" 12-month and possibly "high" 1-month ROR with respect to...*	*1.00; 0.50; 0.00*	*1.00; 1.00; 0.00*	*1.00; 1.00; 0.00*
2	*"Most" of IFs have "low" 12-month and possibly "average" 1-month ROR with respect to...*	0.55; 0.50; 0.23	0.52; 1.00; 0.89	0.30; 1.00; 0.74
3	*"Most" of IFs have "low" 12-month and possibly "low" 1-month ROR with respect to...*	0.75; 0.50; 0.41	0.46; 1.00; 0.71	0.46; 1.00; 0.68

[a] Truth degrees of the linguistic summary ($Z_{\mathbf{Q}}$) and values of quality criteria (10) and (11) computed for the unitary quantifier Q for corresponding W predicates.

incompatible and compatible with the meaning of "most", while all intermediate proportions are treated as compatible to a degree in [0,1].

We focused here on showing the benefits of using contextual and possibly operator in the scope of linguistic data summarization, presenting both a theoretical and semantic justification of this concept and intuitively appealing examples.

Nine linguistic summaries (based on three different triples Q, C, P) reported in Tab. III clearly argue in favour of introduced additional quality criteria (measures) (10)–(11).

First, criterion (10) values 0.5 and 1.0 indicates that all selected contexts are meaningful (see Tab. II).

On the other hand, summaries with highest truth values (all three variants of No. 1 summary) clearly should be discarded — there are no IFs with "high" one-month ROR, which, as we stated at the beginning of section III-C, turns (7) in those summaries into a simple summarizer C (i.e. whole summary into *"Most" of IFs have "low" 12-month ROR.*).

Last three columns of Tab. III confirm that the use of (10) and (11) helps to distinguish interesting summaries (No. 2 and 3 with different W instantiations) from among all with high truth values (rejected summaries are italicized). Additional studies are needed in order to clearly determine the best summaries, yet already the results are promising.

V. CONCLUDING REMARKS

Preliminary computational results of the extension of linguistic data summaries, i.e. bipolar linguistic summaries proposed in [2], demonstrated the need for new quality criteria to determine usefulness of the summary. In [1] we introduced two of them, which have been studied deeper here. The results presented (Tab. III) show that proposed criteria fulfill their role and help select bipolar linguistic summaries valuable and interesting for an end user. Future works in this subject will mainly cover combining introduced criteria with other known quality measures, in order to determine a single value of quality of linguistic summary on one hand, and for evaluating and selecting linguistic summaries by means of heuristic methods, on the other hand.

ACKNOWLEDGMENT

Mateusz Dziedzic contribution is supported by the Foundation for Polish Science under International PhD Projects in Intelligent Computing. Project financed from The European Union within the Innovative Economy Operational Programme (2007-2013) and European Regional Development Fund.

REFERENCES

[1] M. Dziedzic, S. Zadrożny, and J. Kacprzyk, "Bipolar linguistic summaries: a novel fuzzy querying driven approach." in *2013 IFSA-NAFIPS Joint Congress*. Edmonton (Canada): IEEE, 2013, pp. 1279–1284.

[2] ——, "Towards bipolar linguistic summaries: a novel fuzzy bipolar querying based approach." in *Fuzzy Systems (FUZZ-IEEE), 2012 IEEE International Conference on*. Brisbane (Australia): IEEE, 2012, pp. 1–8.

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

[4] J. Kacprzyk, R. R. Yager, and S. Zadrożny, "A fuzzy logic based approach to linguistic summaries of databases," *International Journal of Applied Mathematics and Computer Science*, no. 10, pp. 813–834, 2000.

[5] R. Yager, "On linguistic summaries of data," in *Knowledge Discovery in Databases*, Frawley W. and Piatetsky-Shapiro G., Eds. AAAI/MIT Press, 1991, pp. 347–363.

[6] J. Kacprzyk and S. Zadrożny, "On a fuzzy querying and data mining interface," *Kybernetika*, no. 36, pp. 657–670, 2000.

[7] ——, "Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools," *Inf. Sci.*, vol. 173, no. 4, pp. 281–304, 2005.

[8] S. Zadrożny and J. Kacprzyk, "Bipolar queries: An approach and its various interpretations," in *IFSA/EUSFLAT'09 Conf.*, Lisbon (Portugal), 2009, pp. 1288–1293.

[9] G. Bordogna and G. Pasi, "Linguistic aggregation operators of selection criteria in fuzzy information retrieval," *International Journal of Intelligent Systems*, vol. 10, no. 2, pp. 233–248, 1995.

[10] D. Dubois and H. Prade, "Bipolarity in flexible querying," in *FQAS 2002*, ser. LNAI, T. Andreasen, A. Motro, H. Christiansen, and H. L. Larsen, Eds. Berlin, Heidelberg: Springer-Verlag, 2002, vol. 2522, pp. 174–182.

[11] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," *International Journal of General Systems*, no. 30, pp. 33–154, 2001.

[12] L. Zadeh, "From search engines to question answering systems – the problems of world knowledge relevance deduction and precisiation." in *Fuzzy Logic and the Semantic Web*, E. Sanchez, Ed. Elsevier, 2006, pp. 163–210.

[13] ——, "A computational approach to fuzzy quantifiers in natural languages," *Computers and Mathematics with Applications*, vol. 9, pp. 149–184, 1983.

[14] M. Lacroix and P. Lavency, "Preferences: Putting more knowledge into queries." in *Proceedings of the 13 International Conference on Very Large Databases*, Brighton (UK), 1987, pp. 217–225.

[15] R. Yager, "Higher structures in multi-criteria decision making," *International Journal of Man-Machine Studies*, vol. 36, pp. 553–570, 1992.

[16] ——, "Fuzzy logic in the formulation of decision functions from linguistic specifications," *Kybernetes*, vol. 25, no. 4, pp. 119–130, 1996.

[17] S. Zadrożny and J. Kacprzyk, "Bipolar queries: An aggregation operator focused perspective," *Fuzzy Sets and Systems*, vol. 196, pp. 69–81, 2012.