Bipolar linguistic summaries: a novel fuzzy querying driven approach

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 concept of bipolar linguistic summaries of data, introduced by Dziedzic, Zadrożny and Kacprzyk [1], is further developed. These summaries are meant as an extension of the "classical" linguistic summarization [2], [3], a human-consistent data mining technique, making it possible to express more complex patterns present in data. The focus of the paper is to provide a deeper insight into the very concept of these summaries and illustrating them with some practical examples. Results of preliminary computational experiments are included.**

I. Introduction

Data mining aims at discovering patterns in data that are both interesting to a user and of a practical value. An important aspect is the comprehensiveness of the results presented to a user. A promising way to achieve it is to use (quasi) natural language to express the very essence of the discovered patterns. This has been a motivation for the *linguistic data summaries* by Yager [2] and its further development by Yager [4] and many other contributors, notably Kacprzyk and Zadrożny [5]–[7]. It was quickly observed that linguistic summaries share a lot with *fuzzy linguistic queries* [5] which have been a subject of intensive research and implementations for many decades. While in summaries linguistic terms help to better express discovered regularities in data, in fuzzy linguistic queries they play an analogous role but in expressing user preferences. Thus, such summaries and queries are composed of linguistic terms that can form a user dictionary to be shared by the systems of, respectively, summary generation and query execution. Implementation of both systems "under the same roof" yields a synergy effect as the dictionary of linguistic terms may better reflect their understanding by the user.

Recently, an important role of bipolarity of user preferences is more and more recognized, in particular in fuzzy linguistic querying [8]. Briefly speaking, it boils down to distinguishing positive and negative evaluations which are not necessarily complements of each other. An important line of research focuses here on a special case where negative evaluations are treated as obligatory while the positive evaluations are somehow secondary although this distinction cannot be properly represented by the classical weighting of respective evaluations. This resulted 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 even better model real user preferences exemplified by the query "Find a house, cheap *and possibly* located close to a station".

In our previous paper [1] we started to explore if that intrinsic relation between fuzzy linguistic queries and linguistic data summaries may be adopted for bipolar queries. The answer resulting from our preliminary study was positive and led us to the concept of bipolar linguistic summaries of data. In this paper we further the research in this direction. 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 carried out with various types of queries and corresponding linguistic summaries while Section V discusses the results obtained.

II. An intrinsic relation between the fuzzy linguistic queries and linguistic data summaries

Preferences of a user searching a database may be precise and clear-cut. For example, he or she may be interested in finding all vegetarian restaurants located in a given district of Edmonton. In such a case, classical query languages, such as SQL, provide necessary means to perfectly express these preferences. However, very often the user's preferences are imprecise due to the fact that their original form is a natural language expression. For example, one may be concerned primarily with the cost while looking for an apartment to rent and can express his or her query as:

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

Thus, the user instead of providing a specific price (or, more realistically, an interval of acceptable prices), uses the term "cheap" which is vague but thanks to that directly represents

 DOI: 10.1109/IFSA-NAFIPS.2013.6608586.

the user's expectations. In an approach, referred here to as fuzzy linguistic queries, such terms are represented by fuzzy sets defined in the domains of respective attributes.

The concept of fuzzy linguistic queries has been developed in many papers in which the authors deal with the syntactic and semantic aspects of such queries as well as with effective and efficient methods of their implementation. Usually, a dictionary of linguistic terms is assumed as a part of a system implementing fuzzy linguistic queries. Such a dictionary may contain some predefined linguistic terms and their interpretation in terms of fuzzy sets as well as terms defined by the users, with interpretations possibly evolving in time and/or being context dependent. Linguistic terms collected in a dictionary are a perfect starting point to derive meaningful *linguistic summaries* of a database.

The linguistic summaries are meant as (quasi) natural language expressions that grasp some characteristic features of data collected in a database. The underlying formalism is Zadeh's calculus of linguistically quantified propositions. 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 [2], [4] was the first to propose the use of linguistically quantified propositions to summarize data in a way comprehensible to a user. That idea has been further developed (cf., e.g., Kacprzyk and Yager [11], and Kacprzyk, Yager and Zadrożny [3], [5], [6]). An example of a linguistic summary related to a personnel database may be:

Most employees earn *high* salary, or (2)

Most young employees are *well-educated*. (3)

The following notation is here assumed to formally describe the linguistic summaries and their extensions:

- $R = \{t_1, \ldots, t_n\}$ 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, exemplified by (2) and (3), which is an instantiation of the following abstract *protoform* [12] of type I and type II, respectively:

$$Q_{t \in R}\ S(t) \tag{4}$$

$$Q_{t \in R}\ (U(t), S(t)) \tag{5}$$

The predicates S and U correspond to conditions formed using attributes of the set A. 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";
- a *qualifier* U 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 [2] 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; e.g., $\mu_{most}(0.6) = 0.8$ means that the proportion of 60% of the elements is compatible with the concept of "most" to the degree 0.8. Thus, the truth degrees of the linguistic summaries of type I and II (4)–(5) are, respectively:

$$T(Q_{t \in R} S(t)) = \mu_Q[\frac{1}{n} \sum_{i=1}^{n} \mu_S(t_i)] \tag{6}$$

$$T(Q_{t \in R}(U(t), S(t)) = \mu_Q \left(\frac{\sum_{i=1}^{n} (\mu_U(t_i) \wedge \mu_S(t_i))}{\sum_{i=1}^{n} \mu_U(t_i)} \right). \tag{7}$$

Delgado et al. [14] proposed a set of intuitive properties of models of linguistic quantifiers and showed that some popular approaches might be inadequate. Some approaches are defined only for what they call coherent quantifiers, i.e., which are non-decreasing and $Q(0) = 0, Q(1) = 1$. The only method in their survey, suitable for non-coherent quantifiers, is the one originally introduced by Zadeh, cf. (6)–(7). On the other hand, Zadeh's method is shown to be improper for the traditional crisp quantifiers *exists* and *for all*, and this could clearly be expected.

Delgado et al. [14], [15] proposed a method which starts with the linguistic quantifier Q in the sense of Zadeh and then uses the Choquet integral to evaluate the truth of (4)–(5) as:

$$T(Q_{t \in R} S(t)) = GD_Q(S) = \sum_{i=0}^{n} (b_i - b_{i+1}) \times \mu_Q(i/n), \tag{8}$$

where n is the cardinality of R, b_i is the i-th greatest value of $\mu_S(t_i)$, $i \in \{1, \ldots, n\}$, $b_0 = 1$, $b_{n+1} = 0$, and

$$T(Q_{t \in R}(U(t), S(t))) = GD_Q(S/U) = \sum_{\alpha_i \in M(S/U)} (\alpha_i - \alpha_{i+1}) \times \mu_Q \left(\frac{|(S \cap U)_{\alpha_i}|}{|U_{\alpha_i}|} \right), \tag{9}$$

where $|A|$ denotes the cardinality of the (crisp) set A, $M(S/U)$ is a set of membership degrees α_i such that there exists $t \in R$ and $\mu_{S \cap U}(t) = \alpha_i$ or $\mu_U(t) = \alpha_i$; moreover $1 = \alpha_1 > \alpha_2 \ldots > \alpha_m > \alpha_{m+1} = 0$.

The use of the Choquet integral in such quantifier driven aggregation seems to be promising and we will use $T(Q_{t \in R} S(t)) = GD_Q(S)$ (8) in our computational experiments with bipolar linguistic summaries.

III. Bipolar queries and bipolar linguistic summaries of data

A. Bipolar queries

In classical approaches to preferences modeling, notably in database querying, it is usually assumed that an alternative (tuple) is either accepted or rejected. Thus, it is enough to specify

either positive (acceptance) or negative (rejection) evaluation as the former is just the complement of the latter. 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 queries* against databases where both positive and negative conditions are separately specified [8]. Thus, the evaluation of a bipolar query results in two degrees corresponding to the satisfaction of the positive condition (what is an argument for including a tuple into an answer to the query) and to the satisfaction of the negative condition (what such an inclusion prevents).

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* (a "facultative" desire), denoted P. A subtle relation between conditions C and P may be modeled in different ways [8]. Here we follow the approach of Lacroix and Lavency [16], Yager [17], [18] and Bordogna and Pasi [9], adapted for database querying by Zadrożny and Kacprzyk [19]. We combine both conditions using the "and possibly" operator which aggregates their satisfaction degrees depending on the possibility of a simultaneous matching of both conditions. This possibility is equated with the existence in the database of a tuple which actually satisfies both conditions (to a degree).

Thus, the bipolar query (or, more precisely, its condition), as meant here, may be formally written as:

$$C \text{ and possibly } P, \tag{10}$$

and may be illustrated with the following example of a query:

Find employees that are *young* and possibly earn a *high* salary (11)

We will denote such a bipolar query as (C, P) and interpret it 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 meet both of them. Then the (C, P) query reduces to the conjunction $C \wedge P$. On the other hand, if there is no such a tuple, then it is not possible to satisfy both conditions and P can be ignored. Then the (C, P) query reduces to C. These are however two extreme cases and most interesting is the case when both conditions may be simultaneously satisfied but to a *degree* lower than 1. Then, the matching degree of the (C, P) query against a tuple t belongs to an interval defined by t's matching degrees of $C \wedge P$ and C. This may be formally written as [16]:

$$T\left(C(t) \; and \; possibly \; P(t)\right) = C(t) \wedge \left(\exists s \left(C(s) \wedge P(s)\right) \Rightarrow P(t)\right) \tag{12}$$

The logical connectives used in (12) may be interpreted in various ways within fuzzy logic, to be discussed later on.

B. Bipolar linguistic summaries

In [1] we proposed the concept of a bipolar linguistic summary using as a starting point the concept of a bipolar query and an intrinsic link between fuzzy linguistic queries and "classical" linguistic summaries pointed out earlier in our works. However, the earlier proposed interpretation of "C and possibly P" expressed by (12) makes this proposition true (to a high degree) for a tuple t only if either of two conditions holds: t satisfies both conditions C and P or t satisfies C and there is no tuple in the *whole database* which satisfies both conditions. Thus, the truth of $\exists s C(s) \wedge P(s)$ is fixed for all tuples t and the summarizer C and possibly P does not lead to an interesting linguistic summary; cf. [1]. 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. (13)

The summarizer, which is underlined above, should be meant to be satisfied by an employee 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. (14)

Such a bipolar linguistic summary may now be of interest to the user. Its characteristic feature is the use of a summarizer employing an extended version of the "and possibly" operator, which we will refer to as the "circumstantial and possibly" operator. This operator may be expressed as:

$$C \text{ and possibly } P \text{ with respect to } W. \tag{15}$$

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 (15) is interpreted as:

$$T\left(C(t) \; and \; possibly \; P(t) \; with \; respect \; to \; W\right) = T(C(t) \wedge (\exists s \left(W(t,s) \wedge C(s) \wedge P(s)\right) \Rightarrow P(t))) \tag{16}$$

C. Challenges

There are two main aspects of bipolar linguistic summaries which have to be addressed:

1) interpretation of the logical connectives in (16),
2) quality indices for the bipolar linguistic summaries. (17)

Formula (16) employs the conjunction, disjunction and implication which may be interpreted in many ways within fuzzy logic. We studied this issue for the bipolar queries [8], [19] but

for the context of linguistic summarization and, in particular, in view of the circumstantial bipolarity, the problem needs further studies.

We follow the usual approach of modeling the conjunction and disjunction by the t-norms and t-conorms, respectively [20], which with the corresponding negation operator $\neg$ form the De Morgan triples $(\wedge, \vee, \neg)$. We consider the three De Morgan triples $(\wedge_{\min}, \vee_{\max}, \neg)$, $(\wedge_{\Pi}, \vee_{\Pi}, \neg)$ and $(\wedge_{\mathrm{L}}, \vee_{\mathrm{L}}, \neg)$ (in each case $\neg x = 1 - x$):

t-norms		
MinMax	Minimum	$\min(x, y)$
Π	Product	$x \cdot y$
L	Łukasiewicz	$\max(0, x + y - 1)$
t-conorms		
MinMax	Maximum	$\max(x, y)$
Π	Probabilistic sum	$x + y - x \cdot y$
L	Łukasiewicz	$\min(1, x + y)$

For each De Morgan triple $(\wedge, \vee, \neg)$ we consider two basic implication operators: the S-implication ($x \rightarrow_{S-\vee} y = \neg x \vee y$) and R-implication ($x \rightarrow_{R-\wedge} y = \sup\{z : x \wedge z \leq y\}$) which lead to the following implication operators, respectively:

S-implications		
MinMax	Kleene-Dienes	$\max(1 - x, y)$
Π	Reichenbach	$1 - x + x \cdot y$
L	Łukasiewicz	$\min(1 - x + y, 1)$
R-implications		
MinMax	Gödel	$\begin{cases} 1 & \text{for } x \leq y \\ y & \text{for } x > y \end{cases}$
Π	Goguen	$\begin{cases} 1 & \text{for } x = 0 \\ \min(1, \frac{y}{x}) & \text{for } x \neq 0 \end{cases}$
L	Łukasiewicz	$\min(1 - x + y, 1)$

Computational experiments show that the above De Morgan triples, both with the S- and R-implication, may lead to counter-intuitive results in terms of bipolar queries evaluation.

Thus, in Sections IV-A and IV-B we use the MinMax and Goguen $R-$ implication which turns (16) into:

$$T\left(C(t) \textit{ and possibly } P(t) \textit{ with respect to } W\right) = \begin{cases} \min(C(t), 1) & \text{for } \exists WCP = 0 \\ \min\left(C(t), \min\left(1, \frac{P(t)}{\exists WCP}\right)\right) & \text{otherwise} \end{cases}, \quad (18)$$

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

The second problem mentioned in (17) is related to the very essence of the circumstantial bipolarity concept. Namely, if P and/or W are such that that the premise of the implication in (16) is true to a very low degree *for most* of t's, then the summarizer (15) does not make much sense even if the truth value of a summary involving it is high. Neither such a summarizer make much sense if the premise of the implication in (16) is true to a very high degree *for most* of t's. This is due to the behavior of the bipolar query "C and possibly P" which turn 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, respectively. The introduction of the predicate W partially alleviates this problem, as argued earlier, 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. A solution to this problem may be the use of thresholds on the truth values of the following linguistically quantified propositions:

$$Q_{t \in R} \exists_{s \in R \setminus \{t\}} W(t, s) \quad (19)$$

$$Q_{t \in R} \exists_{s \in R \setminus \{t\}} C(s) \wedge P(s) \wedge W(t, s) \quad (20)$$

as part of the quality evaluation of a bipolar linguistic summary. Namely, if the truth of (19) for a summary is too small (lower than the first threshold value), then such a summary should be discarded. Also, if the truth of (20) 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 is of no interest to the user either. 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 (19)–(20) as if the only tuple related via W with t is only t itself, then the resulting summary is of no interest.

IV. Computational experiments

We present examples of bipolar linguistic summaries and their semantics with respect to their corresponding fuzzy and bipolar queries using data on the rates of return (RORs) of selected investment funds[1] (IFs) – cf. Tab. I.

A. Data querying

In order to compare the semantics of *fuzzy* (1), *standard bipolar* (10) and *circumstantial bipolar* (15) queries, we consider the queries "Show IFs satisfying condition":

(1) **fuzzy**: C *and* P,
(2) **standard bipolar**: C *and possibly* P,
(3) **circumstantial bipolar with crisp cond. *W***: C *and possibly* P with respect to a crisp W,
(4) **circumstantial bipolar with fuzzy cond. *W***: C *and possibly* P with respect to a fuzzy W.

where fuzzy predicates are represented by trapezoidal membership functions defined by four points (x, y, u, w) as shown in Fig. 1. These queries schemes are instantiated as (cf. Fig. 2):

- C: has "high" 12-month ROR,

Table I
Selected investment funds (IF)

#	IF type[a]	1-month ROR[b]	12-month ROR
1	N,S	2.27	39.94
2	N,S	2.47	30.59
3	N,D	3.41	18.14
4	N,D	1.62	17.42
5	F,S	4.29	54.17
6	F,S	2.31	44.22

[a] **N**ational/**F**oreign, **S**tock/**D**ebt securities.
[b] Rate Of Return.

[1] URL: http://www.money.pl/fundusze/ as of April 28, 2013.

Table II
INVESTMENT FUNDS (IF) - TRUTH DEGREES OF FUZZY PREDICATES

#	$C(t)$	$P(t)$	$W(t)$[a]	#:	$W(t,s)$: "same type" 1	2	3	4	5	6	$W(t,s)$: "similar type" 1	2	3	4	5	6
1	1.00	0.27	0.7		1.0	1.0					1.0	1.0			0.5	0.5
2	1.00	0.47	0.7		1.0	1.0					1.0	1.0			0.5	0.5
3	0.00	1.00	0.0				1.0	1.0					1.0	1.0		
4	0.00	0.00	0.0				1.0	1.0					1.0	1.0		
5	1.00	1.00	1.0						1.0	1.0	0.5	0.5			1.0	1.0
6	1.00	0.31	1.0						1.0	1.0	0.5	0.5			1.0	1.0

Empty cell indicates no *similarity* ($W(t,s) = 0.0$).
[a] IF type attribute mapped to $[0,1]$ as: N,S – 0.7, N,D – 0.0, F,S – 1.0.

Table III
INVESTMENT FUNDS – QUERY RESPONSES

#	$T_{\text{query (1)}}$	$T_{\text{query (2)}}$	$T_{\text{query (3)}}$	$T_{\text{query (4)}}$
1	0.27 (*4*)	0.27 (*4*)	0.58 (*3*)	0.54 (*3*)
2	0.47 (*2*)	0.47 (*2*)	1.00 (*1*)	0.94 (*2*)
3	0.00 (*5*)	0.00 (*5*)	0.00 (*5*)	0.00 (*5*)
4	0.00 (*6*)	0.00 (*6*)	0.00 (*6*)	0.00 (*6*)
5	1.00 (*1*)	1.00 (*1*)	1.00 (*2*)	1.00 (*1*)
6	0.31 (*3*)	0.31 (*3*)	0.31 (*4*)	0.31 (*4*)

Tuples ranking is given in parentheses.

- P: has "high" 1-month ROR,
- W_{crisp}: of "the same" type, i.e., $W(t,s)$ is true iff IF type(t) = IF type(s),
- W_{fuzzy}: of "similar" type, defined by $(0.0; 0.0; 0.2; 0.4)$ over $|\text{IF type}(t) - \text{IF type}(s)|$.

Table II shows the truth degrees of these predicates and the mapping of IF type to $[0,1]$ for the tuples from Tab. I; Table III shows the matching degrees of the tuples against queries (1)–(4).

B. Linguistic summaries

Now we present some summaries obtained for data in Tab. I and compare the standard and bipolar linguistic summaries.

In order to facilitate the interpretation we consider only two linguistic quantifiers Q: *majority* (Q_1) and *most* (Q_2). Fig. 3 presents the membership functions of Q_1 and Q_2. The same fuzzy predicates C, P and W from the previous section are used. However, this time C and P are also defined for the linguistic terms "average" and "low". Thus, the summarizers are identical with conditions of the respective queries; e.g., the summary "Majority of IFs have high 12-month and possibly low 1-month ROR with respect to IFs of the same type" (No. 10 in Tab. IV) corresponds to query (3).

The truth values T are evaluated using Zadeh's and Delgado et al.'s approaches, denoted as Z_Q and GD_Q, and defined

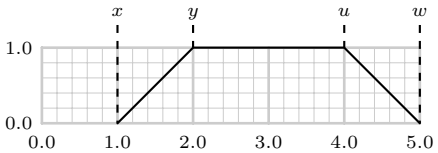

Figure 1. An example of trapezoidal membership function defined by $(1.0, 2.0, 4.0, 5.0)$.

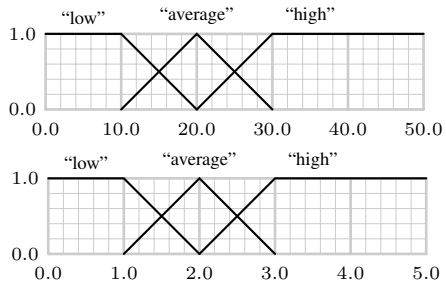

Figure 2. Membership functions of 12-month ROR predicates (condition C – upper plot) and 1-month ROR predicates (condition P – lower plot).

by (6) and (8), respectively, and are shown in Tab. IV. The ranking of the linguistic summaries does not change much for these two approaches: it is the same with respect to the best and worst summaries though there are quantitative differences for the summaries in between.

V. DISCUSSION

We focused on showing the benefits of using circumstantial bipolarity, presenting both a deeper theoretical and semantic justification of this concept and intuitively appealing examples. In the queries, the circumstantial bipolarity employed in queries (3) and (4) manifests itself by involving dynamically for each tuple a context (W) in which the possibility of matching the conditions C and P simultaneously is checked. It may therefore lead to a different ranking of tuples in comparison to a standard bipolar query. In Tab. III it can be seen that the standard bipolar query (2) turns into $C \wedge P$ as tuple #5 fully matches C and P (cf. Tab. II) while in the case of circumstantial bipolar queries (3) the influence of tuple #5 is limited to its context (tuple #6) and, e.g., tuples #1 and #2 are promoted in the ranking. In (4) with its fuzzy condition W, the context of tuples #1 and #2 includes tuple #5 but only to a degree. This slightly lowers the scores for these tuples.

However, the *circumstantial and possibly* operator fully preserves the required and desired character of the conditions C and P as, e.g., tuples #3 and #4 match queries (3)–(4) to degree 0.0 which confirms the fact that a tuple cannot match a query to a degree higher than its matching degree of $C(t)$.

Regarding the bipolar circumstantial linguistic summaries, first, it may be noticed that Zadeh's method is somehow more strict, i.e., gives results closer to 0.0 and 1.0; cf. summaries 1–4 and 8 in Tab. IV. This may justify the use of Delgado et al.'s method which provides for a finer distinction of the

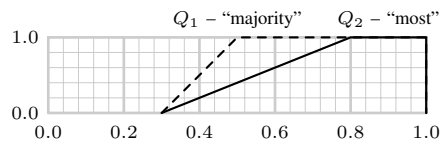

Figure 3. Q_1 ("majority") and Q_2 ("most") linguistic quantifiers' membership functions.

Table IV
EXAMPLES OF LINGUISTIC SUMMARIES (DELGADO ET AL.'S (DG_Q) AND ZADEH'S (Z_Q) APPROACH

#	Linguistic summary[a]	DG_Q	Z_Q	(20)[b]
	Standard linguistic summaries			
1	*"Most" of IF have "high" 12-month and "high" 1-month ROR*	*0.23 (9)*[c]	*0.08 (9)*	
2	*"Most" of IF have "high" 12-month and "average" 1-month ROR*	*0.22 (10)*	*0.05 (10)*	
3	*"Majority" of IF have "high" 12-month and "high" 1-month ROR*	*0.34 (7)*	*0.21 (7)*	
4	**"Majority" of IF have "high" 12-month and "average" 1-month ROR**	**0.56** *(5)*	*0.13 (8)*	
	Bipolar summaries[d]			
5	*"Most" of IF have "high" 12-month and possibly "high" 1-month ROR with respect to IF of "the same" type*	*0.36 (6)*	*0.36 (5)*	*0.42*
6	*"Most" of IF have "high" 12-month and possibly "average" 1-month ROR with respect to IF of "the same" type*	*0.31 (8)*	*0.31 (6)*	*0.44*
7	**"Most" of IF have "high" 12-month and possibly "low" 1-month ROR with respect to IF of "the same" type**	**0.73** *(3)*	**0.73** *(4)*	*0.00*
8	**"Majority" of IF have "high" 12-month and possibly "high" 1-month ROR with respect to IF of "the same" type**	**0.65** *(4)*	**0.90** *(2)*	**0.42**
9	**"Majority" of IF have "high" 12-month and possibly "average" 1-month ROR with respect to IF of "the same" type**	**0.78** *(2)*	**0.77** *(3)*	**0.44**
10	**"Majority" of IF have "high" 12-month and possibly "low" 1-month ROR with respect to IF of "the same" type**	**1.00** *(1)*	**1.00** *(1)*	*0.00*

[a] only summaries with $T > 0.0$ are shown; summaries with $T >= 0.5$ are in **bold**;
[b] quality index (20) computed for the unitary quantifier Q; [c] ranking of tuples is given in the parentheses;
[d] both the crisp and fuzzy conditions W gave similar results; values for the crisp ones are presented.

summaries. However, the ranking obtained for realistic, i.e. sufficiently soft quantifiers, is essentially the same.

Again, by taking into account only a part of the database while evaluating (15) for a tuple we obtain higher truth values of summaries with the circumstantial and possibly operator (cf. summaries No. 4 and 9 in Tab. IV) which makes them "more visible" to the user. Last column of Tab. IV confirms that the use of (20) helps to distinguish interesting summaries (8–9) from among all with high truth values (shown in bold).

VI. CONCLUDING REMARKS

Bipolarity plays an important part in the modeling of user preferences. We provide a further justification that it may also play such a role in data mining while expressing interesting patterns present in datasets. We follow here a path "from queries to summaries" which proved to be fruitful in the case of fuzzy queries and linguistic data summaries. Extensions are needed while involving the bipolarity into what leads to what we call circumstantial bipolar queries and their related bipolar summaries. Some preliminary but promising computational results of the proposed extension strongly argue for its further exploration and development. Future works will mainly cover the examination of different instantiations of the circumstantial and possibly operator, use of other operators (e.g., the standard De Morgan triples, cf. [8], [19]) and the use of other quality criteria for the evaluation of linguistic summaries along the lines of our approach [3].

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, "Towards bipolar linguistic summaries: a novel fuzzy bipolar querying based approach." in *IEEE Int. Conf. on Fuzzy Syst*. Brisbane (Australia): IEEE, 2012, pp. 1–8.

[2] R. Yager, "A new approach to the summarization of data," *Inf. Sci*, vol. 28, pp. 69–86, 1982.

[3] J. Kacprzyk, R. R. Yager, and S. Zadrożny, "A fuzzy logic based approach to linguistic summaries of databases," *Int. J. of Appl. Math. and Comp. Sci*, no. 10, pp. 813–834, 2000.

[4] 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.

[5] J. Kacprzyk and S. Zadrożny, "Data mining via linguistic summaries of databases: an interactive approach," in *A New Paradigm of Knowledge Engineering by Soft Computing*, L. Ding, Ed. Singapore: World Scientific, 2001, pp. 325–345.

[6] ——, "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," *Int. J. of Intell. Syst*, vol. 10, no. 2, pp. 233–248, 1995.

[10] D. Dubois and H. Prade, "Bipolarity in flexible querying," in *FQAS 2002*, Andreasen, T. et al., Ed. Berlin, Heidelberg: Springer-Verlag, 2002, vol. 2522, pp. 174–182.

[11] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," *Int. J. of General Syst*, 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," *Comp. and Math. with Appl*, vol. 9, pp. 149–184, 1983.

[14] M. Delgado, D. Sánchez, and M. A. V. Miranda, "Fuzzy cardinality based evaluation of quantified sentences," *Int. J. of Approx. Reas.*, vol. 23, no. 1, pp. 23–66, 2000.

[15] ——, "A survey of methods for evaluating quantified sentences," in *EUSFLAT-ESTYLF Conf*, Palma de Mallorca, Spain, 1999, pp. 279–282.

[16] M. Lacroix and P. Lavency, "Preferences: Putting more knowledge into queries." in *13th Int. Conf. on Very Large Datab*, Brighton (UK), 1987, pp. 217–225.

[17] R. Yager, "Higher structures in multi-criteria decision making," *Int. J. of Man-Machine Stud*, vol. 36, pp. 553–570, 1992.

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

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

[20] J. Fodor and M. Roubens, *Fuzzy preference modelling and multicriteria decision support*. Kluwer Academic Publishers, 1994.