Linguistic Database Summaries Using Fuzzy Logic: Towards a Human-Consistent Data Mining Tool

Janusz Kacprzyk
Sławomir Zadrożyń
Systems Research Institute, Polish Academy of Sciences

Summary

We discuss an approach to fuzzy linguistic summaries of data (bases) in the sense of Yager, i.e., for instance, if we have a (large) database on employees, and we are interested in a relation between the age and qualifications, then it may be summarized by, say, “most young employees are well qualified”. We present the derivation of such linguistic summaries in the context of Zadeh’s computing with words and perceptions paradigm, and consider his recent idea of a protoform to define and handle more general forms of summaries as proposed by Kacprzyk and Zadrožny. We present an implementation for a small to medium computer retailer, and show how data from the internet can qualitatively enhance the summarization results.

Keywords: decision making, decision support systems, data mining, linguistic summary, natural language generation

1. Introduction

We address the problem of how to deal with too large sets of data that are not comprehensible by the human user. We seek some highly human consistent data mining tools that are simple and possibly inexpensive, preferably based upon the use of (quasi)natural language.

We present an approach advocated in our works (cf. Kacprzyk [4], Kacprzyk, Yager and Zadrożny [7, 8, 9], Kacprzyk and Zadrożny [19, 24, 26]. We propose the use of linguistic data(base) summaries in the sense of Yager (cf. Yager [32], Kacprzyk and Yager [7], Kacprzyk, Yager and Zadrożny [8, 9]) whose essence is that a set of data, e.g. on employees can be summarized linguistically with respect to a an attribute(s) selected by the user or automatically, e.g. age and salaries, by linguistically quantified propositions, e.g. “almost all employees are well qualified”, “most young employees are well paid”, etc. These are simple, human consistent and intuitive statements which are also extremely scalable in a conceptual sense as they are meaningful for all sizes of data sets.

We present the essence of this approach to such summaries, with further extensions (cf. Kacprzyk and Yager [7], Kacprzyk, Yager and Zadrożny [8, 9]) from the perspective of Zadeh’s computing with words and perception paradigm (cf. Zadeh and Kacprzyk [40, 41]), in particular using Zadeh’s [39] concept of a protoform of a fuzzy linguistic summary, adopted to the context of linguistic summaries by Kacprzyk and Zadrożny [24, 26].

We show an implementation of the data summarization in a sales database of a computer retailer and show how this can support decision making.
The basic philosophy of the approach makes use of the computing with words and perception paradigm introduced by Zadeh in the mid-1990s (cf. Zadeh and Kacprzyk’s [40, 41] books). Recently, Zadeh [39] introduced the concept of a protoform, and Kacprzyk and Zadrożny [24, 26] showed that protoforms play a crucial role in the linguistic summarization. One should notice that most perceptions are summaries. For instance, a perception like "most Swedes are tall" is some sort of a summary. It can be represented in Zadeh’s notation as "most As are Bs" which, by a further definition of a protoform as an abstracted summary, may be "QAs are Bs". Notice that we now have a more general, deinstantiated form of our point of departure (most Swedes are tall), and also of "most As are Bs". Most human reasoning is protoform based.

We present an implementation of Yager’s [32, 33] idea, mainly using Kacprzyk and Yager’s [7], and Kacprzyk, Yager and Zadrożny’s [8, 9] extensions. We employ Kacprzyk and Zadrożny’s [13, 21] interactive approach to linguistic summaries via Kacprzyk and Zadrożny’s [10, 11, 12] FQUERY for Access, a fuzzy querying add-on to Access. We show that by relating various types of linguistic summaries to fuzzy queries, with various elements known and sought, we arrive at a hierarchy of protoforms of linguistic summaries which was proposed and extensively studied in Kacprzyk and Zadrożny [24, 26]. Finally, we present an application for a sales database of a computer retailer.

2. Linguistic Data Summaries via Fuzzy Logic with Linguistic Quantifiers

In Yager’s [32, 33] linguistic summary of a set of data, we have: (1) \( V \) - a quality (attribute) of interest, with numeric and non-numeric (e.g. linguistic) values - e.g. salary in a database of workers, (2) \( Y = \{y_1, \ldots, y_n\} \) - a set of objects (records) that manifest quality \( V \), e.g. the set of workers; \( V(y_i) \) - values of quality \( V \) for object \( y_i \), (3) \( D = \{V(y_1), \ldots, V(y_n)\} \) - a set of data (database).

A linguistic summary of data set consists of:
- a summarizer \( S \) (e.g. young),
- a quantity in agreement \( Q \) (e.g. most),
- truth (validity) \( T \) - e.g. 0.7,

as, e.g., "\( T(\text{most of employees are young})=0.7 \)".

The calculation of the truth (validity) of a linguistic summary considered in this section is equivalent to the calculation of the truth value (from [0,1]) of a linguistically quantified statement (e.g., "\( \text{most of the employees are young} \)"), using either Zadeh’s calculus of linguistically quantified statements (cf. Zadeh [37]) or Yager’s OWA operators (cf. Yager and Kacprzyk [35]).

A linguistically quantified proposition, exemplified by "most experts are convinced", is written as "\( Qy \)’s are \( F \)" where \( Q \) is a linguistic quantifier (e.g., most), \( Y = \{y\} \) is a set of objects (e.g., experts), and \( F \) is a (usually) fuzzy property (e.g., convinced). With importance added, \( B \), we get "\( QBy \)’s are \( F \)" , e.g., "most (\( Q \)) of the important (\( B \)) experts (\( y \))’s are convinced (\( F \))". We seek truth( \( Qy \)’s are \( F \) ) or truth( \( QBy \)’s are \( F \)), respectively, knowing truth( \( y \) is \( F \)), \( \forall y \in Y \).
Using Zadeh's [37] fuzzy-logic-based calculus of linguistically quantified propositions, property $F$ and importance $B$ are represented by fuzzy sets in $Y$, and a (proportional, nondecreasing) linguistic quantifier $Q$ is assumed to be a fuzzy set in $[0,1]$ as, e.g.

$$\mu_Q(x) = \begin{cases} 
1 & \text{for } x \geq 0.8 \\
2x - 0.6 & \text{for } 0.3 < x < 0.8 \\
0 & \text{for } x \leq 0.3
\end{cases} \quad (1)$$

Then:

$$\text{truth}(Qy's \text{ are } F) = \mu_Q\left[\frac{1}{n}\sum_{i=1}^{n} \mu_F(y_i)\right] \quad (2)$$
$$\text{truth}(QBy's \text{ are } F) = \mu_Q\left[\frac{\sum_{i=1}^{n} (\mu_B(y_i) \land \mu_F(y_i))}{\sum_{i=1}^{n} \mu_B(y_i)}\right] \quad (3)$$

We can also use an OWA operator (cf. Yager and Kacprzyk [35]) of dimension $p$ which is a mapping $O:[0,1]^p \rightarrow [0,1]$ if with $O$ is associated $W = [w_1, \ldots, w_p]^T$, $w_1 + \cdots + w_p = 1$, $w_i \in [0,1]$, a weighting vector, and ($b_i$ is the $i$-th largest element among $x_1, \ldots, x_p, B = [b_1, \ldots, b_p]$):

$$O(x_1, \ldots, x_p) = w_1 b_1 + \cdots + w_p b_p = W^T B \quad (4)$$

Notice that $w_1 = \cdots = w_{p-1} = 0$ and $w_p = 1$ correspond to the aggregation related to the quantifier “for all” (i.e. taking into account the largest maximal element); $w_1 = 1$ and $w_2 = \cdots = w_p = 0$ correspond to that related to the quantifier “for at least one” (i.e. taking into account the minimal element), through all intermediate situations as, e.g. corresponding to a linguistic quantifier $Q$ when the OWA weights may be found as

$$w_i = \mu_Q\left(\frac{i}{p}\right) - \mu_Q\left(\frac{i-1}{p}\right) \quad (5)$$

For the OWA operators for importance qualified data, suppose that we have $A = [a_1, \ldots, a_p]$, and a vector of importances $V = [v_1, \ldots, v_p]$ such that $v_i \in [0,1]$ is the importance of $a_i, i = 1, \ldots, p$, $v_1 + \cdots + v_p = 1$. Then (4) becomes:

$$O_V (x_1, \ldots, x_p) = \bar{w}_1 b_1 + \cdots + \bar{w}_p b_p = \bar{W}^T B \quad (6)$$

where
\[ w_j = \mu_Q \left( \frac{\sum_{k=1}^{j} u_k}{\sum_{k=1}^{p} u_k} \right) - \mu_Q \left( \frac{\sum_{k=1}^{j-1} u_k}{\sum_{k=1}^{p} u_k} \right) \] (7)

in which \( u_k \) is the importance of \( b_k \), the \( k \)-largest element of \( A \) (i.e. the corresponding \( v_k \)).

3. A Fuzzy Querying Add-on for Formulating Linguistic Summaries

In Kacprzyk and Zadroży’s [13, 21] approach, it is assumed that in practice a fully automatic determination of an aspect that is interesting for the user is impossible, and interactivity, i.e. user assistance, is in the definition of summarizers (indication of attributes and their combinations) necessary. This is via a user interface of a fuzzy querying add-on, FQUERY for Access which has been proposed and developed by Kacprzyk and Zadroży [10, 11, 12] inspired by Kacprzyk and Ziółkowski’s [28] and Kacprzyk, Zadroży and Ziółkowski’s [27] proposals of database queries with fuzzy linguistic quantifiers. Obviously, fuzzy queries directly correspond to summarizers in linguistic summaries. Thus, the derivation of a linguistic summary may proceed in an interactive (user-assisted) way as follows: (1) the user formulates a set of linguistic summaries of interest (relevance) using the fuzzy querying add-on, (2) the system retrieves records from the database and calculates the validity of each summary adopted, and (3) a most appropriate linguistic summary is chosen.

We restate linguistic summarization in the fuzzy querying context. The two propositions with fuzzy linguistic quantifiers (without and with importance) are now:

"Most records match query \( S \)" (8)

"Most records meeting conditions \( B \) match query \( S \)" (9)

In database terminology, \( B \) corresponds to a filter and (9) claims that most records passing through \( B \) match query \( S \).

The concept of a protoform (Zadeh [39]) is highly relevant here. A protoform is defined as an abstract prototype, that is, for the query (summary) given by (8) and (9) as, respectively:

"Most \( R \)'s are \( S \)" (10)

"Most \( BR \)'s are \( S \)" (11)

where \( R \) means "records", \( B \) is a filter, and \( S \) is a query.

Evidently, protoforms may form a hierarchy, and we can define higher level (more abstract) protoforms, e.g. replacing most by a general linguistic quantifier \( Q \):

"\( QR \)'s are \( S \)" (12)

"\( QBR \)'s are \( S \)" (13)

Basically, the more abstract forms correspond to cases in which we assume less about summaries sought. In Table 1 five basic types of linguistic summaries are shown.
Table 1: Classification of linguistic summaries

<table>
<thead>
<tr>
<th>Type</th>
<th>Given</th>
<th>Sought</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S</td>
<td>Q</td>
<td>Simple summaries through ad-hoc queries</td>
</tr>
<tr>
<td>2</td>
<td>S B</td>
<td>Q</td>
<td>Conditional summaries through ad-hoc queries</td>
</tr>
<tr>
<td>3</td>
<td>Q S\textsuperscript{structure}</td>
<td>S\textsuperscript{value}</td>
<td>Simple value oriented summaries</td>
</tr>
<tr>
<td>4</td>
<td>Q S\textsuperscript{structure} B</td>
<td>S\textsuperscript{value}</td>
<td>Conditional value oriented summaries</td>
</tr>
<tr>
<td>5</td>
<td>Nothing</td>
<td>S B Q</td>
<td>General fuzzy rules</td>
</tr>
</tbody>
</table>

where $S^{\text{structure}}$ denotes that attributes and their connection in a summary are known, while $S^{\text{value}}$ denotes that these values are sought; some placeholders for numerical values are left.

Type 1 may be easily obtained by a simple extension of FQUERY. The user has to construct a query – a candidate summary, and it is to be determined which fraction of rows matches that query and which linguistic quantifier best denotes this fraction. Type 2 is a straightforward extension of Type 1 by adding a fuzzy filter. The summaries of Type 1 and 3 have been implemented as an extension to FQUERY for Access. For an extensive analysis of various issues related to the use of protoforms in the context of summarization, see Kacprzyk and Zadrozny [24, 26].

4. Implementation

Our implementation concerns a computer retailer, and we deal with its sales database whose basic structure is shown in Table 2.

Table 2. The basic structure of the database (in the "dbf" type format)

<table>
<thead>
<tr>
<th>Attribute name</th>
<th>Attribute type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Date</td>
<td>Date of sale</td>
</tr>
<tr>
<td>Time</td>
<td>Time</td>
<td>Time of sale transaction</td>
</tr>
<tr>
<td>Name</td>
<td>Text</td>
<td>Name of the product</td>
</tr>
<tr>
<td>Amount (number)</td>
<td>Numeric</td>
<td>Number of products sold in the transaction</td>
</tr>
<tr>
<td>Price</td>
<td>Numeric</td>
<td>Unit price</td>
</tr>
<tr>
<td>Commission</td>
<td>Numeric</td>
<td>Commission (in %) on sale</td>
</tr>
<tr>
<td>Value</td>
<td>Numeric</td>
<td>Value = amount (number) x price; of the product</td>
</tr>
<tr>
<td>Discount</td>
<td>Numeric</td>
<td>Discount (in %) for transaction</td>
</tr>
<tr>
<td>Group</td>
<td>Text</td>
<td>Product group to which the product belongs</td>
</tr>
<tr>
<td>Transaction value</td>
<td>Numeric</td>
<td>Value of the whole transaction</td>
</tr>
<tr>
<td>Total sale to customer</td>
<td>Numeric</td>
<td>Total value of sales to the customer in fiscal year</td>
</tr>
<tr>
<td>Purchasing frequency</td>
<td>Numeric</td>
<td>Number of purchases by customer in fiscal year</td>
</tr>
<tr>
<td>Town</td>
<td>Test</td>
<td>Town where the customer lives or is based</td>
</tr>
</tbody>
</table>

We will now give some examples of linguistic summaries. First, suppose that we seek a relation between the commission and the type of goods sold. The best linguistic summaries obtained are as in Table 3.
Table 3. Linguistic summaries expressing relations between the group of products and commission

<table>
<thead>
<tr>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>About ½ of sales of network elements is with a high commission</td>
</tr>
<tr>
<td>About ½ of sales of computers is with a medium commission</td>
</tr>
<tr>
<td>Much sales of accessories is with a high commission.</td>
</tr>
<tr>
<td>Much sales of components is with a low commission.</td>
</tr>
<tr>
<td>About ½ of sales of software is with a low commission.</td>
</tr>
<tr>
<td>About ½ of sales of computers is with a low commission.</td>
</tr>
<tr>
<td>A few sales of components is without commission.</td>
</tr>
<tr>
<td>A few sales of computers is with a high commission.</td>
</tr>
<tr>
<td>Very few sales of printers is with a high commission.</td>
</tr>
</tbody>
</table>

In Table 4 we show some linguistic summaries expressing relations between the attributes: size of customer, regularity of customer (purchasing frequency), date of sale, time of sale, commission, group of product and day of sale. This is an example of the most sophisticated form of linguistic summaries supported by the system described.

Table 4. Linguistic summaries expressing relations between the attributes: size of customer, regularity of customer (purchasing frequency), date of sale, time of sale, commission, group of product and day of sale

<table>
<thead>
<tr>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Much sales on Saturday is about noon with a low commission</td>
</tr>
<tr>
<td>Much sales on Saturday is about noon for bigger customers</td>
</tr>
<tr>
<td>Much sales on Saturday is about noon</td>
</tr>
<tr>
<td>Much sales on Saturday is about noon for regular customers</td>
</tr>
<tr>
<td>A few sales for regular customers is with a low commission</td>
</tr>
<tr>
<td>A few sales for small customers is with a low commission</td>
</tr>
<tr>
<td>A few sales for one-time customers is with a low commission</td>
</tr>
<tr>
<td>Much sales for small customers is for nonregular customers</td>
</tr>
</tbody>
</table>

Notice that the linguistic summaries obtained provide much of relevant and useful information, and can help the decision maker make decisions.

Recently (cf. Kacprzyk and Zadrożyń [26]) we extended the above system. Notice that the above summaries involve only data from the own database. However, no company operates in a vacuum, and some external data (e.g. on climate when the operation and/or results depend on climatic conditions, national and global economic indicators, etc.) should be taken into account. In our system, climatic data are accounted for. Such data are widely available and the Internet is the best source of such data, inexpensive and easily available.

Now, relations between group of products, time of sale, temperature, precipitation, and type of customers, the best linguistic summaries (of both our “internal” data from the sales database, and “external” meteorological data from an Internet service) are as in Table 5.
Table 5. Linguistic summaries expressing relations between the attributes: group of products, time of sale, temperature, precipitation, and type of customers

<table>
<thead>
<tr>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very few sales of software in hot days to individual customers</td>
</tr>
<tr>
<td>About 1/2 of sales of accessories in rainy days on weekends by the end of the year</td>
</tr>
<tr>
<td>About 1/3 of sales of computers in rainy days to individual customers</td>
</tr>
</tbody>
</table>

Notice that the use of external data gives a new quality to possible linguistic summaries.

5. Conclusion

We presented an interactive, fuzzy logic based approach to linguistic summarization of databases, human consistent and easily comprehensible by human beings. Through the use of Zadeh’s computing with words and perceptions paradigm, and of protoforms we attained a high human consistency and scalability. The implementation presented shows that linguistic summaries can be helpful for real life decision making in a company.

Bibliography


Janusz Kacprzyk
Sławomir Zadrożyński
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warszawa
e-mail: kacprzyk@ibspan.waw.pl
zadrozy@ibspan.waw.pl