Chapter 10
Protoforms of Linguistic Database Summaries as a Human Consistent Tool for Using Natural Language in Data Mining

Janusz Kacprzyk
Polish Academy of Sciences, Poland

Sławomir Zadrożny
Polish Academy of Sciences, Poland

ABSTRACT

We consider linguistic database summaries in the sense of Yager (1982), in an implementable form proposed by Kacprzyk & Yager (2001) and Kacprzyk, Yager & Zadrożny (2000), exemplified by, for a personnel database, "most employees are young and well paid" (with some degree of truth) and their extensions as a very general tool for a human consistent summarization of large data sets. We advocate the use of the concept of a protoform (prototypical form), vividly advocated by Zadeh and shown by Kacprzyk & Zadrożny (2005) as a general form of a linguistic data summary. Then, we present an extension of our interactive approach to fuzzy linguistic summaries, based on fuzzy logic and fuzzy database queries with linguistic quantifiers. We show how fuzzy queries are related to linguistic summaries, and that one can introduce a hierarchy of protoforms, or abstract summaries in the sense of latest Zadeh's (2002) ideas meant mainly for increasing deduction capabilities of search engines. We show an implementation for the summarization of Web server logs.

DOI: 10.4018/978-1-4666-0261-8.ch010

INTRODUCTION

Data summarization is one of basic capabilities needed by any "intelligent" system. Since for the human being the only fully natural means of communication is natural language, a linguistic summarization would be very desirable, exemplified by, for a data set on employees, a statement (linguistic summary) "almost all young and well qualified employees are well paid".

This may clearly be an instance of a paradigm shift that is advocated in recent time whose prominent example is the so-called "computing with words (and perceptions) paradigm" introduced by Zadeh in the mid-1990s, and extensively presented in Zadeh & Kacprzyk's (1999) books.

Unfortunately, data summarization is still in general unsolved a problem. Very many techniques are available but they are not "intelligent enough", and not human-consistent, partly due to a limited use of natural language.

We show here the use of linguistic database summaries introduced by Yager (1982, 1991, 1995, 1996), and then considerably advanced by Kacprzyk (2000), Kacprzyk & Yager (2001), and Kacprzyk, Yager & Zadrożny (2000, 2001), Zadrożny & Kacprzyk (1999), and implemented in Kacprzyk & Zadrożny (1998,2000a-d, 2001a-e, 2002, 2003, 2005). We derive here linguistic data summaries as linguistically quantified propositions as, e.g., "most of the employees are young and well paid", with a degree of truth (validity), in case of a personnel database.

We employ Kacprzyk & Zadrożny's (1998, 2000a-d, 2001) interactive approach to linguistic summaries in which the determination of a class of summaries of interest is done via Kacprzyk & Zadrożny's (1994, 1995a-b, 2001b) FQUERY for Access, a fuzzy querying add-in to Microsoft Access, extended to the querying over the Internet in Kacprzyk & Zadrożny (2000b). Since a fully automatic generation of linguistic summaries is not feasible at present, an interaction with the user is assumed for the determination of a class of summaries of interest, and this is done via the above fuzzy querying add-in.

Extending Kacprzyk & Zadrożny (2002), we show that by relating various types of linguistic summaries to fuzzy queries, with various known and sought elements, we can arrive at a hierarchy of prototypical forms, or – in Zadeh's (2002) terminology – protoforms, of linguistic data summaries. This seems to be a very powerful conceptual idea.

We present an implementation of the proposed approach to the derivation of linguistic summaries for Web server logs. This implementation may be viewed as a step towards the implementation of protoforms of linguistic summaries.

LINGUISTIC SUMMARIES USING FUZZY LOGIC WITH LINGUISTIC QUANTIFIERS

In Yager's (1982) approach, we have:

- V is a quality (attribute) of interest, e.g. salary in a database of workers,
- $Y = \{y_1, ..., y_n\}$ is a set of objects (records) that manifest quality V, e.g. the set of workers; hence $V(y_i)$ are values of quality V for object y_i,
- $D = \{V(y_1), ..., V(y_n)\}$ is a set of data (the "database" on question)

A *linguistic summary* of a data set (data base) consists of:

- a summarizer S (e.g. young),
- a quantity in agreement Q (e.g. most),
- truth T - e.g. 0.7,
- a qualifier R (optionally), i.e. another linguistic term (e.g. well-earning), determining a fuzzy subset of Y.

as, e.g., "T(*most* of employees are *young*)=0.7". The truth T may be meant more generally as, e.g., validity.

Given a set of data D, we can hypothetize any appropriate summarizer S and any quantity in