40

Computing with Words and Protoforms: Powerful and Far Reaching Ideas

Janusz Kacprzyk and Sławomir Zadrożny

Abstract. We show how Zadeh's computing with words and perceptions, the idea of an extraordinary power and far reaching impact, can lead to a new direction in the use of natural language in data mining, the linguistic data(base) summaries. We emphasize the relevance of Zadeh's protoform which may effectively and efficiently represent the user's intentions and interests, and show that various types of linguistic data summaries may be viewed as items in a hierarchy of protoforms of summaries.

40.1 Introduction

We wish to shortly present the essence and some applications of *computing with words* (CWW), and its inherent *protoforms*. These can be considered, in our opinion, to be the most influential and far reaching idea conceived by Zadeh, except for his "grand inventions" like fuzzy sets and possibility theories or foundations of the state space approach in systems modeling. To follow the spirit of this volume, our exposition will be concise and comprehensible.

Computing with words (and perceptions), or CWW, introduced by Zadeh in the mid-90s, and first comprehensively presented in Zadeh and Kacprzyk's books [17], may be viewed a new "technology" in the representation, processing and solving of various real life human centric problems. It makes it possible to use natural language, with its inherent imprecision, in an effective and efficient way.

Zadeh used the so-called PNL (precisiated natural language) in which statements about values, relations between variables, etc. are represented by constraints. Its statements, written "x isr R", may be different, and correspond to numeric values, intervals, possibility, verity and probability distributions, usuality qualification, rough sets representations, fuzzy relations, etc. For us, the usuality qualified statements have been be of special relevance. Basically, it says "x is usually R" that is meant as "in most cases, x is R". PNL may play various roles among which crucial are: description of perceptions, definition of sophisticated concepts, a language for perception based reasoning, etc. Notice that the usuality is an example of a modality in natural language. Clearly, this all is meant as a tool for the representation and processing of perceptions.

Another Zadeh's ingenious inception is the concept of a *protoform* [16]. In general, most perceptions are summaries, exemplified by "most Swedes are tall" which

R. Seising et al. (Eds.): *On Fuzziness: Volume 1*, STUDFUZZ 298, pp. 265–270.
DOI: 10.1007/978-3-642-35641-4_40

is clearly a summary of the Swedes with respect to height. It can be represented in Zadeh's notation as "most *A*s are *B*s". This can be employed for reasoning under various assumptions. One can go a step further, and define a protoform as an abstracted summary, like "*QA*s are *B*s", and now have a more general, deinstantiated form of our point of departure (most Swedes are tall), and also of "most *A*s are *B*s". Most of human reasoning is protoform based.

Basically, the essence of our work over the years was to show that the concept of PNL, and in particular of a protoform, viewed from the perspective of CWW, can be of use in attempts at a more effective and efficient use of vast information resources, notably through linguistic data(base) summaries which are very characteristic for human needs and comprehension abilities. In what follows we give an outline of our approach.

40.2 Linguistic Data Summaries via Fuzzy Logic with Linguistic Quantifiers

The linguistic summary is meant as a sentence [in a (quasi)natural language] that subsumes the very essence (from a certain point of view) of a set of data. Here this set is assumed to be numeric, large and not comprehensible in its original form by the human being. In Yager's approach (cf. Yager [12], Kacprzyk and Yager [3], and Kacprzyk, Yager and Zadrożny [4]), if $Y = \{y_1, \ldots, y_n\}$ is a set of records in a database, e.g., representing the set of workers, and $A = \{A_1, \ldots, A_m\}$ is a set of attributes characterizing the elements of Y, e.g., salary, age, etc., $A_j(y_i)$ denotes a value of A_j for object y_i, then a *linguistic summary* of a data set Y consists of: (1) a summarizer S, i.e. an attribute together with a linguistic green (fuzzy predicate) (e.g. "low salary" for attribute "salary"), (2) a quantity in agreement Q, i.e. a linguistic quantifier (e.g. most), and (3) truth (validity) T of the summary, $T \in [0,1]$; optionally, a qualifier R, i.e. another attribute together with a linguistic term (fuzzy predicate) may be added (e.g. "young" for "age").

The linguistic summaries, without and with a qualifier, may be exemplified by

$$T(\text{most of employees earn low salary}) = 0.7 \tag{40.1}$$

$$T(\text{most of young employees earn low salary}) = 0.85 \tag{40.2}$$

The core of a linguistic summary is a *linguistically quantified proposition* in the sense of Zadeh [15]; those corresponding to (40.1) and (40.2) may be written, respectively, as

$$Qy\text{'s are } S \tag{40.3}$$

$$QRy\text{'s are } S \tag{40.4}$$

The T, i.e., the truth value of (40.3) or (40.4), can be calculated by using either original Zadeh's calculus of linguistically quantified statements (cf. [15]), or other interpretations of linguistic quantifiers.

Formulas (40.3) and (40.4) may be seen as the most abstract protoforms, the highest in the hierarchy of protoforms, while (40.1) and (40.2) are examples of fully instantiated protoforms, "leaves" of their "hierarchy tree". Going down this hierarchy one has to instantiate particular components of (40.3) and (40.4), i.e., quantifier Q and fuzzy predicates S and R. The instantiation of the former one boils down to the selection of a quantifier. The instantiation of fuzzy predicates requires the choice of attributes together with linguistic terms and a structure they form when combined using logical connectives. Thus, in general, there is an infinite number of potential protoforms, though, due to a limited capability of the user only a reasonable number of summaries should be taken into account.

The concept of a protoform may provide a guiding paradigm for the design of a user interface supporting the mining of linguistic summaries. It may be assumed that the user specifies a protoform of linguistic summaries sought. Basically, the more abstract protoform the less should be assumed about summaries sought, i.e., the wider range of summaries is expected by the user. There are two limit cases, where:

- a totally abstract protoform is specified, i.e., (40.4),
- all elements of a protoform are totally specified as given linguistic terms.

In the former case the system has to construct all possible summaries for the context of a given database and show those with the highest validity T. In the second case, the whole summary is specified by the user and the system has only to verify its validity. The former case is usually more attractive for the user but more complex computationally. There is a number of intermediate cases that may be more practical. In Table 40.1 basic types of protoforms/linguistic summaries are shown, corresponding to protoforms of a more and more abstract form.

Table 40.1. Classification of protoforms/linguistic summaries

Type	Protoform	Given	Sought
0	QRy's are S	Everything	validity T
1	Qy's are S	S	Q
2	QRy's are S	S and R	Q
3	Qy's are S	Q and structure of S	linguistic terms in S
4	QRy's are S	Q, R and structure of S	linguistic terms in S
5	QRy's are S	Nothing	S, R and Q

Basically, each of fuzzy predicates S and R may be defined by listing its atomic fuzzy predicates (i.e., pairs of "attribute/linguistic term") and structure, i.e., how these atomic predicates are combined. In Table 40.1 S (or R) corresponds to the full description of both the atomic fuzzy predicates as well as the structure. For example: "*Q young* employees earn a *high salary*" is a protoform of Type 2, while "Most employees earn a *"?" salary*" is a protoform of Type 3. In the first case the

system has to select a linguistic quantifier for which the proposition is true (valid) to a high degree. In the second case, the linguistic quantifier and (only) the structure of summarizer S are given and the system has to choose a linguistic term to replace the question mark ("?") yielding a highly valid proposition.

Thus, the use of protoforms makes it possible to devise a uniform procedure to handle a wide class of linguistic data summaries so that the system can be easily adaptable to a variety of situations, users' interests and preferences, scales of the project, etc.

An interesting extension of the concept of a linguistic summary to the linguistic summarization of time series data was shown in a series of works by Kacprzyk, Wilbik and Zadrożny [1, 2]. In this case the array of possible protoforms is much larger as it reflects various perspectives, intentions, etc. of the user. The protoforms used in those works may be exemplifed by: "Among all y's, Q are P", which may be instantiated as "Among all segments (of the time series) most are slowly increasing", and "Among all R segments, Q are P", which may be instantiated as "Among all short segments almost all are quickly decreasing", as well as more sophisticated protoforms, for instance temporal ones like: "E_T among all y's Q are P", which may be instantiated as "Recently, among all segments, most are slowly increasing", and "E_T among all Ry's Q are P", which may be instantiated as "Initially, among all short segments, most are quickly decreasing"; they both go beyond the classic Zadeh's protoforms.

It is easy to notice that the mining of linguistic summaries may be viewed to be closely related to *natural language generation* (NLG) and this path was suggested in Kacprzyk and Zadrożny [11]. This may be a promising direction as NLG is a well developed area and software is available.

40.3 Mining of Linguistic Data Summaries

In Kacprzyk and Zadrożny's [9] interactive approach, the mining of summaries proceeds via a user interface of a fuzzy querying add-on such as FQUERY for Access [5, 6, 10]. In such an add-on a dictionary of linguistic terms is maintained, such as "'young", "most" etc. This terms are then readily available as building blocks of a summary.

Thus, the derivation of a linguistic summary of Type 0 in Table 40.1 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) the most appropriate (highly valid) linguistic summaries are chosen.

Referring to Table 40.1, we can observe that the summaries of Type 1-4 may be produced by a simple extension of such a querying add-on as FQUERY for Access. On the other hand, the discovery of general Type 5 rules, which may be equated with the fuzzy IF-THEN rules, is difficult, and some simplifications about the structure

of fuzzy predicates and/or quantifier are needed. Kacprzyk and Zadrożny [7, 8] proposed to distinguish a subclass of Type 5 summaries which may be interpreted as *fuzzy association rules* and mined using adapted versions of well-known algorithms, e.g., Apriori.

40.4 Concluding Remarks

We show how Zadeh's ingenious idea of computing with words and perceptions, based on his concept of a precisiated natural language (PNL), can lead to a new direction in the use of natural language in data mining, the linguistic data(base) summaries. We emphasize the relevance of Zadeh's protoform, and show that various types of linguistic data summaries may be viewed as items in the hierarchy of protoforms of linguistic summaries.

Acknowledgement. To Professor Lotfi Zadeh who – through his ingenious idea of computing with words and protoforms – has provided all of us with tools for an effective and efficient use of natural language in a vast array of systems modeling, data mining, knowledge discovery, ... tasks.

References

1. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic Summarization of Time Series Using a Fuzzy Quantifier Driven Aggregation. Fuzzy Sets and Systems 159, 1485–1499 (2008)
2. Kacprzyk, J., Wilbik, A., Zadrożny, S.: An Approach to the Linguistic Summarization of Time Series Using a Fuzzy Quantifier Driven Aggregation. International Journal of Intelligent Systems 25, 411–439 (2010)
3. Kacprzyk, J., Yager, R.R.: Linguistic Summaries of Data Using Fuzzy Logic. International Journal of General Systems 30, 33–154 (2001)
4. Kacprzyk, J., Yager, R.R., Zadrożny, S.: A Fuzzy Logic Based Approach to Linguistic Summaries of Databases. International Journal of Applied Mathematics and Computer Science 10, 813–834 (2000)
5. Kacprzyk, J., Zadrożny, S.: FQUERY for Access: Fuzzy Querying for a Windows-based DBMS. In: Bosc, P., Kacprzyk, J. (eds.) Fuzziness in Database Management Systems, pp. 415–433. Springer, Heidelberg (1995)
6. Kacprzyk, J., Zadrożny, S.: The Paradigm of Computing with Words in Intelligent Database Querying. In: Zadeh, L.A., Kacprzyk, J. (eds.) Computing with Words in Information/Intelligent Systems. Part 2. Foundations, pp. 382–398. Springer, Heidelberg (1999)
7. Kacprzyk, J., Zadrożny, S.: On Combining Intelligent Querying and Data Mining Using Fuzzy Logic Concepts. In: Bordogna, G., Pasi, G. (eds.) Recent Research Issues on the Management of Fuzziness in Databases, pp. 67–81. Springer, Heidelberg (2000)
8. Kacprzyk, J., Zadrożny, S.: Computing with Words: Towards a New Generation of Linguistic Querying and Summarization of Databases. In: Sinčak, P., Vaščak, J. (eds.) Quo Vadis Computational Intelligence?, pp. 144–175. Springer, Heidelberg (2000)

9. Kacprzyk, J., Zadrożny, S.: Data Mining via Linguistic Summaries of Databases: An Interactive Approach. In: Ding, L. (ed.) A New Paradigm of Knowledge Engineering by Soft Computing, pp. 325–345. World Scientific, Singapore (2001)
10. Kacprzyk, J., Zadrożny, S.: Computing with Words in Intelligent Database Querying: Standalone and Internet-based Applications. Information Sciences 134, 71–109 (2001)
11. Kacprzyk, J., Zadrożny, S.: Computing with Words is an Implementable Paradigm: Fuzzy Queries, Linguistic Data Summaries, and Natural Language Generation. IEEE Transactions on Fuzzy Systems 18, 461–472 (2010)
12. Yager, R.R.: A New Approach to the Summarization of Data. Information Sciences 28, 69–86 (1982)
13. Yager, R.R., Kacprzyk, J. (eds.): The Ordered Weighted Averaging Operators: Theory and Applications. Kluwer, Boston (1997)
14. Yager, R.R., Kacprzyk, J., Beliakov, G. (eds.): Recent Developments in the Ordered Weighted Averaging Operators: Theory and Practice. Springer, Heidelberg (2011)
15. Zadeh, L.A.: A Computational Approach to Fuzzy Quantifiers in Natural Languages. Computers and Mathematics with Applications 9, 149–184 (1983)
16. Zadeh, L.A.: A Prototype-centered Approach to Adding Deduction Capabilities to Search Engines - The Concept of a Protoform, BISC Seminar. University of California, Berkeley (2002)
17. Zadeh, L.A., Kacprzyk, J. (eds.): Computing with Words in Information/Intelligent Systems, Part 1. Foundations. Part 2. Applications. Springer, Heidelberg (1999)