Linguistic Summarization of Trends: A Fuzzy Logic Based Approach

Janusz Kacprzyk[1,2], Anna Wilbik[1] and Sławomir Zadrożny[1]

[1] Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
[2]Warsaw Information Technology (WIT)
ul. Newelska 6, 01-447 Warsaw, Poland
{kacprzyk, wilbik, zadrozny}@ibspan.waw.pl

Abstract

The purpose of this paper is to propose a new easily implementable approach to a linguistic summarization of trends that may be present in temporal data. It is based on fuzzy linguistic summaries of data (databases) in the sense of Yager (cf. Yager [20], Kacprzyk and Yager [9], and Kacprzyk, Yager and Zadrożny [10]) which in the form of natural language-like sentences subsume the very essence of a set of data.

Keywords: linguistic summary, trend analysis, fuzzy logic, computing with words.

1 Introduction

Trend analysis is very important and widely used in many fields, like climatology or economics (e.g., [6, 8, 18]). Recently many methods of trend analysis have been introduced (e.g., [3, 5, 7, 15, 16]). Most of them are based on the detection of change of the first and second derivative.

Linguistic summaries of databases [20, 9] might be seen as having a similar goal in a more general context.

Such summaries are meant to capture essential features of original data according to the user's needs. Thus, they provide the user with a simpler view of data collected in a database. Therefore some attempts have been made to apply them to temporal data, too (cf., e.g., [4]).

Generally, data summarization is still unsolved a problem in spite of vast research efforts. Very many techniques are available but they are neither "intelligent enough", nor human consistent, partly due to the fact that the use of natural language is limited. This concerns, e.g., summarizing statistics, exemplified by the average, median, minimum, maximum, α-percentile, etc. which – in spite of recent efforts to soften them – are still far from being able to reflect a real human perception of essence of their extended meaning.

In this paper we propose a new method of an effective and efficient trend summarization.

2 Temporal data and trend analysis

Temporal data is meant as numerical data that changes over time. Time series is a sequence of data items measured typically at uniformly spaced time moments. A time window is a common technique used for trend analysis in time series data. A time window is defined as a set of observations in a chronological order that describe a specified sequence of continuous observations over a time span specified by this time window [17].

In this paper we will identify trends as linear increasing or decreasing functions. For this purpose we will represent given time series data as a piece-wise linear function. In particular, we will use the concept of a uniform partially linear approximation of time series.

A function f is a uniform ε-approximation of a time series or a set of points $\{(x_i, y_i)\}$ if for a given, context dependent $\varepsilon > 0$, there holds

$$\forall i: \ |f(x_i) - y_i| \leq \varepsilon \tag{1}$$

Sklansky and Gonzalez [19] have proposed an efficient and interesting algorithm that finds the linear uniform ε-approximation for subsets of points. The algorithm constructs the intersection of cones starting from a point p_i of the time series and including the circle of radius ε around the subsequent data points p_{i+j}, $j = 1, \ldots$ until the intersection of all cones starting at p_i is empty. If for p_{i+k} the intersection is empty, then we construct a new cone starting at p_{i+k-1}. Figure 1 presents the idea of the algorithm. The family of possible solutions is indicated with a gray area.

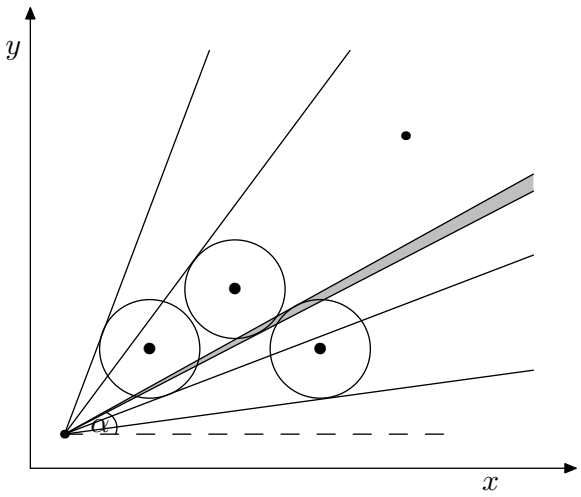

Figure 1: An illustration of the algorithm [19] for uniform ε-approximation

As a result it may be taken either a single line, chosen as, e.g. a bisector, or one that minimizes the distance (e.g. assumed as sum of squared errors, SSE) from the approximated points, or the whole family of possible solutions.

This method is effective and fast as it requires only a single pass through the data.

3 Trend description

While summarizing trends we should consider the following three factors:

- dynamics
- duration
- variability

In what follows we will briefly discuss these factors.

3.1 Dynamics

Under the term *dynamics* we understand the speed of changes. It can be described by the slope of a line representing the trend, (cf. α in Fig. 1). Thus, to quantify dynamics we may use the interval of possible angles $\alpha \in \langle -90; 90 \rangle$.

However it might be impractical to use such a scale directly while describing trends. Therefore we may use a fuzzy granulation in order to meet the users' needs and task specificity. The user may construct a scale of linguistic terms corresponding to various directions of a trend line as, e.g.:

- quickly decreasing
- decreasing
- slowly decreasing
- constant
- slowly increasing
- increasing
- quickly increasing

Figure 2 illustrates the lines corresponding to the particular linguistic terms. In fact, each term represents a fuzzy granule of directions. In [1, 2] there are presented many methods of constructing such a fuzzy granulation. The user may define a membership functions of particular linguistic terms depending on his needs; for example as shown in Figure 3.

We map a single value α (or the whole interval of the angles corresponding to the gray area in Figure 2) characterizing the dynamics of a trend identified using Sklansky and Gonzalez method [19], into a fuzzy set best matching

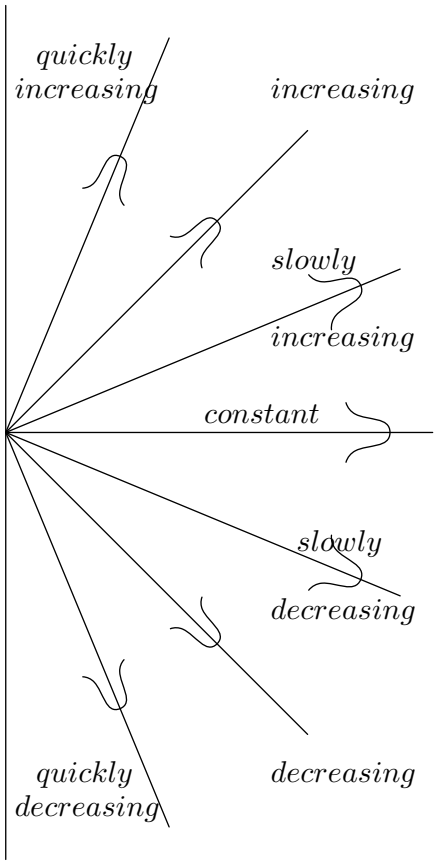

Figure 2: A visual representation of angle granules defining dynamics

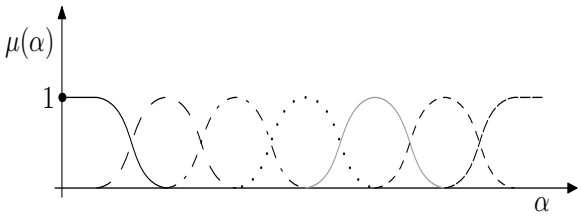

Figure 3: Example of membership functions of the linguistic terms describing the dynamics of the trends

given angle. Then we will say that a given trend is, e.g., "decreasing to a degree 0.8", if $\mu_{decreasing}(\alpha) = 0.8$, where $\mu_{decreasing}$ is the membership function of a fuzzy set representing the linguistic term "decreasing" that is a best match for the angle α characterizing the trend under consideration.

3.2 Duration

Duration describes the length of a single trend found by the linear uniform ε-approximation. Again we will treat it as a linguistic variable. An example of its linguistic labels is "long trend" defined as a fuzzy set, whose membership function might be assumed as in Figure 4, where OX is the axis of time measured with units that are used in the time series data under consideration.

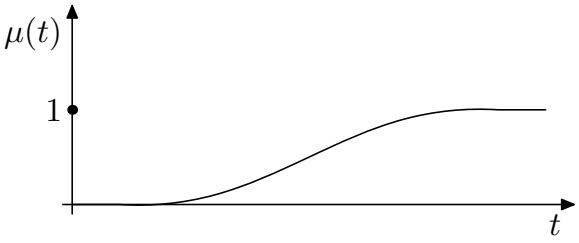

Figure 4: Example of membership function describing the term "long" concerning the trend duration

The actual definitions of linguistic terms describing the duration depend on the perspective assumed by the user. He or she, analyzing the data, may adopt this or another time horizon implied by his or her needs. The analysis may be a part of a policy, strategic or tactical planning, and thus, may require a global or local look, respectively.

3.3 Variability

Variability refers to how "spread out" (in the sense of values taken on) a group of data is. There are five frequently used statistical measures of variability:

- the range (maximum - minimum)
- the interquartile range (IQR) calculated as the third quartile[1] minus the first quartile[2] that may be interpreted as representing the middle 50% of the data.
- the variance is calculated as
$$\frac{\sum_i (x_i - \bar{x})^2}{n},$$
where $\bar{x}$ is the mean value.
- the standard deviation – a square root of the variance

[1] third quartile is the 75th percentile
[2] first quartile is the 25th percentile

- the mean absolute deviation (MAD), calculated as

$$\frac{\sum_i |x_i - \bar{x}|}{n}$$

We propose to measure the variability of a trend as the distance of the data points covered by this trend from their linear uniform ε-approximation (cf. Section 3) that represents given trend. For this purpose we propose to employ a distance between the point and the family of possible solutions, indicated as a gray cone in Figure 1. Equation (1) assures that the distance is definitely smaller than ε. We may use this information for the normalization. The normalized distance equals 0 if the point lays in the gray area. In the opposite case it is equal to the distance of the nearest point belonging to the cone, divided by ε.

Alternatively, we may bisect the cone and then compute the distance between the point and this ray.

In both cases we can use the sign as an indicator of the side at which the point lies. Again the measure of variability is treated as a linguistic variable and expressed using linguistic terms (labels) modeled by fuzzy sets defined by the user.

4 Linguistic summaries

A linguistic summary, as presented in [14, 13] is meant as a natural language-like sentence that subsumes the very essence of a set of data. This set is assumed to be numeric and is usually large, not comprehensible in its original form by the human being. In Yager's approach (cf. Yager [20], Kacprzyk and Yager [9], and Kacprzyk, Yager and Zadrożny [10]) the following context for linguistic summaries mining is assumed:

- $Y = \{y_1, \ldots, y_n\}$ is a set of objects (records) in a database, e.g., the set of workers;
- $A = \{A_1, \ldots, A_m\}$ is a set of attributes characterizing objects from Y, e.g., salary, age, etc. in a database of workers, and $A_j(y_i)$ denotes a value of attribute A_j for object y_i.

A linguistic summary of a data set D consists of:

- a summarizer S, i.e. an attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute A_j (e.g. "low salary" for attribute "salary");
- a quantity in agreement Q, i.e. a linguistic quantifier (e.g. most);
- truth (validity) T of the summary, i.e. a number from the interval $[0, 1]$ assessing the truth (validity) of the summary (e.g. 0.7); usually, only summaries with a high value of T are interesting;
- optionally, a qualifier R, i.e. another attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute A_k determining a (fuzzy subset) of Y (e.g. "young" for attribute "age").

Thus, a linguistic summary may be exemplified by

$$T(\text{most of employees earn low salary}) = 0.7 \tag{2}$$

A richer form of a linguistic summary may include a qualifier (e.g. young) as in, e.g.,

$$T(\text{most of young employees earn low salary}) = 0.7 \tag{3}$$

Thus, basically, the core of a linguistic summary is a *linguistically quantified proposition* in the sense of Zadeh [21]. A linguistically quantified proposition, corresponding to (2) may be written as

$$Qy\text{'s are } S \tag{4}$$

and the one corresponding to (3) may be written as

$$QRy\text{'s are } S \tag{5}$$

Then, the component of a linguistic summary, T, i.e., its truth (validity), directly corresponds to the truth value of (4) or (5). This may be calculated by using either original Zadeh's calculus of linguistically quantified propositions (cf. [21]), or other interpretations of linguistic quantifiers.

5 Trend summarization

In order to characterize the summaries of trends we will refer to Zadeh's concept of a protoform (cf., Zadeh [22]). Basically, a protoform is defined as a more or less abstract prototype (template) of a linguistically quantified proposition. Then, summaries mentioned above might be represented by protoforms of the following form:

- We may consider a short form of summaries:

  $$Q \text{ trends are } S \tag{6}$$

  like e.g.,

  > *Most* of trends have *a large variability*

- We may also consider an extended form of the summary represented by the following protoform:

  $$QR \text{ trends are } S \tag{7}$$

  and exemplified by:

  > *Most* of *slowly decreasing trends* have *a large variability*

Using Zadeh's [21] fuzzy logic based calculus of linguistically quantified propositions, a (proportional, nondecreasing) linguistic quantifier Q is assumed to be a fuzzy set in the interval $[0, 1]$ as, e.g.

$$\mu_Q(x) = \begin{cases} 1 & \text{for } x \geq 0.8 \\ 2x - 0.6 & \text{for } 0.3 < x < 0.8 \\ 0 & \text{for } x \leq 0.3 \end{cases} \tag{8}$$

The truth values (from [0,1]) of (6) and (7) are calculated, respectively as

$$\text{truth}(Qy\text{'s are } S) = \mu_Q \left[\frac{1}{n} \sum_{i=1}^{n} \mu_S(y_i) \right] \tag{9}$$

and

$$\text{truth}(QRy\text{'s are } S) = \\ = \mu_Q \left[\frac{\sum_{i=1}^{n} (\mu_R(y_i) \wedge \mu_S(y_i))}{\sum_{i=1}^{n} \mu_R(y_i)} \right] \tag{10}$$

Both the fuzzy predicates S and R are assumed above to be of a rather simplified, atomic form referring to just one attribute. They can be extended to cover more sophisticated summaries involving some confluence of various attribute values as, e.g, "slowly decreasing and short".

6 Example

Let us assume that we have discovered from some given data trends listed in Table 1. We assume the granulation of dynamics presented in Section 3.1.

Table 1: Example of trends.

id	dynamics (α)	duration (time units)	variability ([0,1])
1	25	15	0.2
2	-45	1	0.3
3	75	2	0.8
4	-40	1	0.1
5	-55	1	0.7
6	50	2	0.3
7	-52	1	0.5
8	-37	2	0.9
9	15	5	0.0

We can consider the following trend summarization:

> *Most* of trends are *decreasing*

In this summary *most* is the linguistic quantifier Q. The membership function is as in (8).

"The trends are decreasing" is a summarizer S with the membership function of the "decreasing" term given in (11).

$$\mu_S(x) = \begin{cases} 0 & \text{for } \alpha \leq -65 \\ 0,066\alpha + 4.333 & \text{for } -65 < \alpha < \\ & < -50 \\ 1 & \text{for } -50 \leq \alpha \leq \\ & \leq -40 \\ -0.05\alpha - 1 & \text{for } -40 < \alpha < \\ & < -20 \\ 0 & \text{for } \alpha \geq -20 \end{cases} \tag{11}$$

n is the number of all trends. In this example n=9.

The truth is computed according to the formula (9):

$$\text{truth}(\textit{Most} \text{ of the trends are } \textit{decreasing}) = \\ = \mu_Q \left[\frac{1}{n} \sum_{i=1}^{n} \mu_S(y_i) \right] = 0.389$$

If we consider an extended form, we may have the following summary:

Most of *short* trends are *decreasing*

Again, *most* is the linguistic quantifier Q with the membership function defined as (8). *"The trends are decreasing"* is a summarizer S as in the previous example. *"The trend is short"* is the qualifier R. We define the membership function μ_R as follows:

$$\mu_R(x) = \begin{cases} 1 & \text{for } t \leq 1 \\ -\frac{1}{2}t + \frac{3}{2} & \text{for } 1 < t < 3 \\ 0 & \text{for } t \geq 3 \end{cases}$$

The truth is computed according to the formula (10):

$$\text{truth}(\textit{Most} \text{ of } \textit{short} \text{ trends are } \textit{decreasing}) = \\ = \mu_Q \left[\frac{\sum_{i=1}^{n} (\mu_R(y_i) \wedge \mu_S(y_i))}{\sum_{i=1}^{n} \mu_R(y_i)} \right] = \\ = 0.892$$

These summaries are based on the frequency of the trends of given type identified in a data set. Thus, many short trends can influence or even dominate long, although rare trends. A further research is needed to solve this problem.

7 Concluding remarks

In the paper we have proposed an effective and efficient method for temporal data summarization. We identify linear trends in data using the algorithm proposed in [19]. Then, having the trends spotted, we propose to characterize them with 3 basic attributes concerning their dynamics, variability and duration. Finally, the data set obtained is described with a set of linguistic summaries in the sense of Yager.

We plan to integrate the generation of linguistic summaries concerning time series data with our FQUERY for Access system developed for the fuzzy queries (cf., [11, 12]). Thus, further research will focus on the search for other possible ways/attributes to characterize the trends as well as on the implementation of the method.

References

[1] I. Batyrshin. On granular derivatives and the solution of a granular initial value problem. *International Journal Applied Mathematics and Computer Science*, 12(3):403–410, 2002.

[2] I. Batyrshin and L. Sheremetov. Perception based functions in qualitative forecasting. to appear.

[3] I. Batyrshin, L. Sheremetov, and R. Herrera-Avelar. Perception based patterns in time series data mining. to appear.

[4] D.-A. Chiang, L. R. Chow, and Y.-F. Wang. Mining time series data by a fuzzy linguistic summary system. *Fuzzy sets and Systems*, 112:419–432, 2000.

[5] J. Colomer, J. Melendez, J. L. de la Rosa, and J. Augilar. A qualitative/quantitative representation of signals for supervision of continuous systems. In *Proceedings of the European Control Conference -ECC97*, Brussels, 1997.

[6] H. Feidas, T. Makrogiannis, and E. Bora-Senta. Trend analysis of air temperature time series in greece and their relationship with circulation using surface and satellite data: 1955-2001. *Theoretical and Applied Climatology*, 79(3-4):185–208, 2004.

[7] F. Hoeppner. *Knowledge Discovery from Sequential Data*. PhD thesis, Braunschweig University, 2003.

[8] B. G. Hunt. A description of persistent climatic anomalies in a 1000-year climatic model simulation. *Climate Dynamics*, 17(9):717–733, 2001.

[9] J. Kacprzyk and R. Yager. Linguistic summaries of data using fuzzy logic. *International Journal of General Systems*, 30:33–154, 2001.

[10] J. Kacprzyk, R. Yager, and S. Zadrożny. A fuzzy logic based approach to linguistic summaries of databases. *International Journal of Applied Mathematics and Computer Science*, 10:813–834, 2000.

[11] J. Kacprzyk and S. Zadrożny. FQUERY for access: fuzzy querying for a windows-based dbms. In P. Bosc and J. Kacprzyk, editors, *Fuzziness in Database Management Systems*, pages 415–433. Springer-Verlag, Heidelberg, 1995.

[12] J. Kacprzyk and S. Zadrożny. The paradigm of computing with words in intelligent database querying. In L. Zadeh and J. Kacprzyk, editors, *Computing with Words in Information/Intelligent Systems. Part 2. Foundations*, pages 382–398. Springer–Verlag, Heidelberg and New York, 1999.

[13] J. Kacprzyk and S. Zadrożny. Fuzzy linguistic data summaries as a human consistent, user adaptable solution to data mining. In B. Gabrys, K. Leiviska, and J. Strackeljan, editors, *Do Smart Adaptive Systems Exist?*, pages 321–339. Springer, Berlin, Heidelberg, New York, 2005.

[14] J. Kacprzyk and S. Zadrożny. Linguistic database summaries and their protoforms: toward natural language based knowledge discovery tools. *Information Sciences*, 173:281–304, 2005.

[15] S. Kivikunnas. Overview of process trend analysis methods and applications. In *Proceedings of Workshop on Applications in Chemical and Biochemical Industry*, Aachen, 1999.

[16] K. B. Konstantinov and T. Yoshida. Real-time qualitative analysis of the temporal shapes of (bio) process variables. *American Institute of Chemical Engineers Journal*, 38(11):1703–1715, 1992.

[17] A. Kusiak and A. Burns. Mining temporal data: A coal-fired boiler case study. In *Knowledge Based Intelligent Information and Engineering Systems*, LNAI 3683, pages 953–958. 2005.

[18] K. Neusser. An investigation into a nonlinear stochastic trend model. *Empirical Economics*, 24(1):135–153, 1999.

[19] J. Sklansky and V. Gonzalez. Fast polygonal approximation of digitized curves. *Pattern Recognition*, 12(5):327–331, 1980.

[20] R. Yager. A new approach to the summarization of data. *Information Sciences*, 28:69–86, 1982.

[21] L. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers and Mathematics with Applications*, 9:149–184, 1983.

[22] L. Zadeh. A prototype-centered approach to adding deduction capabilities to search engines – the concept of a protoform. In *BISC Seminar*, University of California, Berkeley, 2002.