

Linguistic Summaries of Time Series via a Quantifier Based Aggregation Using the Sugeno Integral

Janusz Kacprzyk, *Fellow, IEEE*, Anna Wilbik, and Sławomir Zadrozny

Abstract—Linguistic summaries as descriptions of trends in time series data are proposed. Two general types of such summaries are discussed. The point of departure are linguistic summaries of databases due to Yager. The specificity of time series summarization requires a more general approach to linguistic quantifier based aggregation. Sugeno integrals are employed to address this problem.

I. INTRODUCTION

Linguistic summary is meant as a concise, human-consistent description of a data set. The very concept has been introduced by Yager [13] and further developed by Kacprzyk and Yager [8], and Kacprzyk, Yager and Zadrozny [9]. In this approach the content of a database is summarized via a natural language like expression, semantics of which is provided in the framework of the Zadeh's calculus of the linguistically quantified propositions [14].

In this paper we consider a specific type of data, namely time series, i.e. a certain real valued function of time. For a manager, stock exchange players etc. it might be worthwhile to obtain a brief, natural language like description of trends present in the data on a company performance, stock exchange quotations etc. over a certain period of time. This is not meant as a replacement for a classical statistical analysis but rather as an additional form of data description characterized by its high human consistency.

The summaries we propose refer to trends identified here with straight line segments of the piece-wise linear approximation of time series. Thus, the first step is the construction of such an approximation. For this purpose we use a modified version of the simple, easy to use Sklansky and Gonzalez algorithm presented in [12]. Then we employ a set of features (attributes) to characterize the trends such as the slope of the line, the fairness of approximation of the original data points by the line segment and length of the period of time comprising the trend.

Basically the summaries proposed by Yager are interpreted in terms of the number or proportion of elements possessing a certain property. In the framework considered here a summary might look like: "Most of the trends are short" or in a more sophisticated form: "Most long trends are increasing". Such expressions are easily interpreted using Zadeh's calculus of the linguistically quantified propositions. The

most important element of this interpretation is a linguistic quantifier exemplified by "most". In Zadeh's approach it is interpreted in terms of a proportion of elements possessing a certain property (e.g., a length of a trend) among all the elements considered (e.g., all trends). In [7] we have proposed to use Yager's linguistic summaries, interpreted in the framework of Zadeh's calculus, for time series.

Another type of summaries we propose here do not use the linguistic quantifier based aggregation over the number of trends but over the time instants they take altogether. For example, an interesting summary may take the following form: "Trends taking most of time are increasing" or "Increasing trends taking most of the time are of a low variability". Such summaries do not directly fit the framework of the original Yager's approach. In order to overcome this difficulty we generalize our previous approach (cf. Kacprzyk, Wilbik and Zadrozny [7]), modelling the linguistic quantifier based aggregation both over the number of trends as well over the time they take with the use of the Sugeno integral.

The paper is organized as follows. First we describe the way the trends are extracted from time series and characterized using a set of attributes. Then we briefly remind the basics of the original Yager's approach to linguistic summarization and discuss how it may be used to describe a set of trends. In the next section we show how these summaries might be interpreted using the concept of fuzzy measure and the Sugeno integral. Finally we present some simple examples of linguistic summaries of an artificial data set.

II. CHARACTERIZATION OF TIME SERIES

In our approach time series data $\{(x_i, y_i)\}$ are approximated by a piece-wise linear function f such that for a given $\varepsilon > 0$, there holds

$$\forall i : |f(x_i) - y_i| \leq \varepsilon \quad (1)$$

There exist many algorithms that find such approximations (cf. [5], [6]). Our starting point is the Sklansky and Gonzalez algorithm [12] that seems to be a good choice due to its simplicity and efficiency. We modified it in the following way. 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} , $j = 1, \dots$ until this intersection becomes empty. If for p_{i+k} the intersection is empty, then the points $p_i, p_{i+1}, \dots, p_{i+k-1}$ are approximated by a straight line segment and to approximate the remaining points we construct a new cone starting at p_{i+k-1} . Figure 1 presents the idea of the algorithm. The

Janusz Kacprzyk is with Systems Research Institute, Polish Academy of Sciences ul. Newelska 6, 01-447 Warsaw, Poland and Warsaw Information Technology (WIT) ul. Newelska 6, 01-447 Warsaw, Poland e-mail: kacprzyk@ibspan.waw.pl

Anna Wilbik and Sławomir Zadrozny are with Systems Research Institute, Polish Academy of Sciences ul. Newelska 6, 01-447 Warsaw, e mail: {wilbik, zadrozny}@ibspan.waw.pl

family of possible solutions, i.e., straight line segments to approximate points p_1 and p_2 , is indicated with a dark gray area.

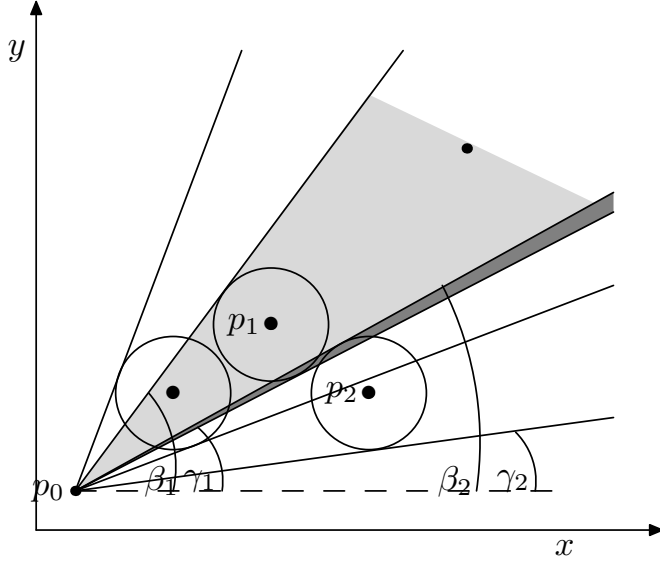


Fig. 1. An illustration of the algorithm [12] for an uniform ε -approximation

To make it more intuitively appealing we will now present the algorithm in the form of a pseudocode. First denote by:

- p_0 – the current starting point,
- p_1 – the last point successfully checked, i.e. for it the intersection of the cones starting at p_0 is non-empty,
- p_2 – the next point to be checked
- Alpha_01 – a pair of angles (γ_1, β_1) , meant as an interval that defines the current cone, as shown in Figure 1 (indicated by light gray and dark gray area)
- Alpha_02 – a pair of angles defining the cone constructed to check p_2 (i.e., the cone starting at point p_0 and inscribing the circle of radius ε around the point p_2 (cf. (γ_2, β_2) in Figure 1))
- function $\text{read_point}()$ fetches the next data point,
- function $\text{find}()$ finds a pair of angles defining the cone starting at point p_0 and inscribing the circle of radius ε around of the point p_2

The pseudocode of the procedure that extracts the trends is depicted in Figure 2.

The bounding values of Alpha_02 (γ_2, β_2) , computed by function $\text{find}()$, are the slopes of two lines such that:

- they are tangent to the circle of radius ε around point p_2
- they start at the point p_0

Let $\Delta x = x_0 - x_2$ and $\Delta y = y_0 - y_2$ then the angles γ_2, β_2 can be expressed by the formulas:

$$\gamma_2 = \arctg \left(\frac{(\Delta x)(\Delta y) - \varepsilon \sqrt{1 - (\Delta x)^2 - (\Delta y)^2}}{(\Delta x)^2 - \varepsilon^2} \right)$$

$$\beta_2 = \arctg \left(\frac{(\Delta x)(\Delta y) + \varepsilon \sqrt{1 - (\Delta x)^2 - (\Delta y)^2}}{(\Delta x)^2 - \varepsilon^2} \right)$$

```

read_point(p_0);
read_point(p_1);
do
{
  p_2 = p_1;
  Alpha_02 = find();
  Alpha_01 = Alpha_02;
  do
  {
    Alpha_01 = Alpha_01 ∩ Alpha_02;

    p_1=p_2;
    read_point(p_2);
    Alpha_02 = find();
  } while(Alpha_01 ∩ Alpha_02 ≠ ∅);
  save_found_trend();
  p_0 = p_1;
  p_1 = p_2;
}

```

Fig. 2. Pseudocode of the procedure for extracting trends. For technical reasons we do not use Greek letters to denote variables.

Then, as an approximation of points p_0, \dots, p_1 we assume either a single straight line segment, 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, i.e. the segments of the rays of the cone.

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

We characterize the trends, meant as the straight line segments of the above described uniform ε -approximation, using the following three features:

- dynamics of change
- duration
- variability

In what follows we will briefly discuss these factors.

A. Dynamics of change

Under the term *dynamics of change* we understand the speed of changes. It can be described by the slope of a line representing the trend, (cf. any angle η from the interval $\langle \gamma, \beta \rangle$ in Figure 1). Thus, to quantify dynamics of change we may use the interval of possible angles $\eta \in \langle -90; 90 \rangle$ or their trigonometrical transformation.

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 3 illustrates the lines corresponding to the particular linguistic terms. In fact, each term represents a fuzzy granule

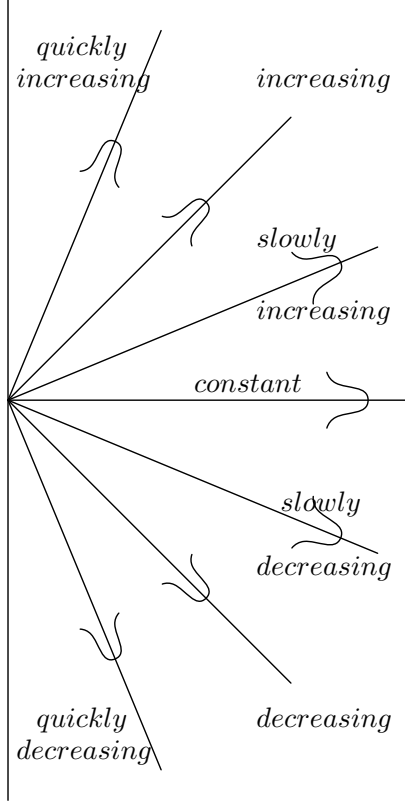


Fig. 3. A visual representation of angle granules defining dynamics of change

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 or her needs.

We map a single value η (or the whole interval of the angles corresponding to the gray area in Figure 1) characterizing the dynamics of change of a trend identified using the algorithm shown in Figure 2, into a fuzzy set best matching a given angle. Then we will say that a given trend is, e.g., “decreasing to a degree 0.8”, if $\mu_{decreasing}(\eta) = 0.8$, where $\mu_{decreasing}$ is the membership function of a fuzzy set representing the linguistic term “decreasing” that is a best match for the angle η characterizing the trend under consideration.

B. Duration

Duration describes the length of a single trend. Again we will treat it as a linguistic variable. An example of its linguistic labels is “long” 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.

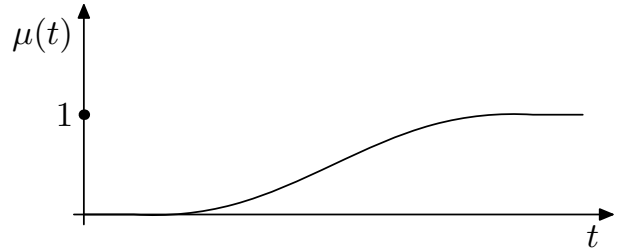


Fig. 4. Example of membership function describing the term “long” concerning the trend duration

The actual definitions of linguistic terms describing the duration depends 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.

C. 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). Although this range is computationally the easiest measure of variability, it is not widely used as it is only based on two extreme data points. This make it very vulnerable to outliers and therefore may not adequately describe real variability.
- the interquartile range (IQR) calculated as the third quartile¹ minus the first quartile² that may be interpreted as representing the middle 50% of the data. It is resistant to outliers and is computationally as easy as the range.
- the variance is calculated as $1/n \sum_i (x_i - \bar{x})^2$, where \bar{x} is the mean value.
- the standard deviation – a square root of the variance. Both the variance and the standard deviation are affected by extreme values.
- the mean absolute deviation (MAD), calculated as $1/n \sum_i |x_i - \bar{x}|$. While it has a natural intuitive definition as the “mean deviation from the mean”, the introduction of the absolute value makes analytical calculations using this statistics much more complicated.

We propose to measure the variability of a trend as the distance of data points covered by this trend from a linear uniform ε -approximation that represents a given trend. For this purpose we propose to employ a distance between a point and a family of possible solutions, indicated as a dark 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 dark gray area. In the opposite case it is

¹third quartile is the 75th percentile

²first quartile is the 25th percentile

equal to the distance to 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.

Again the measure of variability is treated as a linguistic variable whose values are linguistic terms (labels) modeled by fuzzy sets defined by the user.

III. LINGUISTIC SUMMARIES AND THEIR APPLICATION TO TREND SUMMARIZATION

A linguistic summary, as presented in [10], [11] 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 [13], Kacprzyk and Yager [8], and Kacprzyk, Yager and Zadrozny [9]) the following context for linguistic summaries mining is assumed:

- $Y = \{y_1, \dots, y_n\}$ is a set of objects (records) in a database, e.g., the set of workers;
- $A = \{A_1, \dots, A_m\}$ 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 P , 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, meant as a proposition of Zadeh's calculus of linguistically quantified propositions (e.g. 0.7), 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. i.e. another attribute A_k 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").

In what follows we will often for brevity identify summarizers and qualifiers with the linguistic terms they contain. In particular we will refer to the membership function μ_P or μ_R of the summarizer or qualifier to be meant as the membership functions of respective linguistic terms.

Thus, a linguistic summary may be exemplified by

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

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

$$T(\text{most of young employees earn low salary}) = 0.9 \quad (3)$$

Thus, basically, the core of a linguistic summary is a *linguistically quantified proposition* in the sense of Zadeh

[14]. A linguistically quantified proposition corresponding to (2) may be written as

$$Qy's \text{ are } P \quad (4)$$

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

$$QRy's \text{ are } P \quad (5)$$

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 the original Zadeh's calculus of linguistically quantified propositions (cf. [14]), or via other interpretations of linguistic quantifiers. The truth values (from $[0, 1]$) of (4) and (5) are calculated, respectively, as

$$T(Qy's \text{ are } P) = \mu_Q \left(\frac{1}{n} \sum_{i=1}^n \mu_P(y_i) \right) \quad (6)$$

$$T(QRy's \text{ are } P) = \mu_Q \left(\frac{\sum_{i=1}^n (\mu_R(y_i) \wedge \mu_P(y_i))}{\sum_{i=1}^n \mu_R(y_i)} \right) \quad (7)$$

where Q is a fuzzy set representing the linguistic quantifier in the sense of Zadeh [14].

In order to characterize the summaries of trends we will refer to Zadeh's concept of a protoform (cf., Zadeh [15]). 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 two types of the protoforms of the following forms. We may consider *frequency based summaries* and we obtain:

- a simple form:

$$Q \text{ trends are } P \quad (8)$$

exemplified by:

Most of trends have a large variability

- an extended form:

$$QR \text{ trends are } P \quad (9)$$

exemplified by:

Most of slowly decreasing trends have a large variability

However it should be noticed that in some cases the summaries of the above types might not properly grasp the character of time series. For example, assuming there are many very short trends of high variability and a few long terms of very low variability we may obtain a summary stating that "Most of trends have a large variability". This might be perceived as somehow incomplete or inaccurate summarization on its own as in fact the trends taking *most of time* have very low variability. Thus we propose to complement the above types of summaries with two more types of *duration based summaries*. These may be represented by the following schemes:

- a simple form:

$$\text{The trends that took } Q \text{ time are } P \quad (10)$$

exemplified by:

The trends that took *most* time have a *large variability*

- an extended form:

$$R \text{ trends that took } Q \text{ time are } P \quad (11)$$

exemplified by:

Slowly decreasing trends that took *most* time have a *large variability*

The truth degrees T of the frequency based summaries (8)-(9) can be directly computed using Zadeh's calculus of linguistically quantified propositions, in particular the formulae (6) and (7) are of use. However this is not the case when we consider duration based summaries. The reason is that in case of (8)-(9) a linguistic quantifier aggregates over the number of trends possessing a certain property while in case of (10)-(11) this aggregation goes over time taken by the trends. Thus, in the former case the *count* (number) of the trends matters that is properly accounted for with the use of the Σ -Count cardinality in formulae (6) and (7). In the latter case however another mode of aggregation is required. In order to secure a unified solution in both cases we propose to employ the Sugeno integral as explained in the next section.

IV. LINGUISTIC SUMMARY INTERPRETATION VIA THE SUGENO INTEGRAL

As we explained in the previous section, duration based linguistic summaries do not fit well to the interpretation of a linguistically quantified proposition employed in Zadeh's calculus. Thus we propose here to use the Sugeno integral for that purpose.

Let us start with a brief recall of the basics of the Sugeno integral. Let $X = \{x_1, \dots, x_n\}$ be a finite set. Then, (cf., e.g., [4]) a *fuzzy measure* on X is a set function $\mu : \mathcal{P}(X) \rightarrow [0,1]$ such that:

$$\begin{aligned} \mu(\emptyset) &= 0, \mu(X) = 1 \\ \text{if } A \subseteq B \text{ then } \mu(A) &\leq \mu(B), \forall A, B \in \mathcal{P}(X) \end{aligned} \quad (12)$$

where $\mathcal{P}(X)$ denotes a set of all subsets of X .

Let μ is a fuzzy measure on X . The *discrete Sugeno integral* of function $f : X \rightarrow [0,1]$, $f(x_i) = a_i$, with respect to μ is a function $S_\mu : [0,1]^n \rightarrow [0,1]$ such that

$$S_\mu(a_1, \dots, a_n) = \max_{i=1, \dots, n} (a_{\sigma(i)} \wedge \mu(B_i)) \quad (13)$$

where \wedge stands for the minimum, σ is such a permutation of $\{1, \dots, n\}$ that $a_{\sigma(i)}$ is the i -th smallest element from among the a_i 's and $B_i = \{x_{\sigma(i)}, \dots, x_{\sigma(n)}\}$.

We can treat function f as a membership function of a fuzzy set $F \in \mathcal{F}(X)$, where $\mathcal{F}(X)$ denotes a family of

fuzzy sets defined in X . Then the Sugeno integral can be equivalently defined as a function $S_\mu : \mathcal{F}(X) \rightarrow [0,1]$ such that

$$S_\mu(F) = \max_{\alpha_i \in \{a_1, \dots, a_n\}} (\alpha_i \wedge \mu(F_{\alpha_i})) \quad (14)$$

where F_{α_i} is the α -cut of F and the meaning of other symbols is as in (13).

The fuzzy measure and the Sugeno integral may be intuitively interpreted in the context of multicriteria decision making (MCDM) where we have a set of criteria and some options (decisions) characterized by the degree of satisfaction of particular criteria. In such a setting X is a set of criteria and μ expresses the importance of each subset of criteria, i.e., how the satisfaction of a given subset of criteria contributes to the overall evaluation of the option. Then the properties of the fuzzy measure (12) properly require that the satisfaction of all criteria makes an option fully satisfactory and that the more criteria are satisfied by an option the better its overall evaluation. Finally the set F represents an option and $\mu_F(x)$ defines the degree to which it satisfies the criterion x . Then the Sugeno integral may be interpreted as an aggregation operator yielding an overall evaluation of option F in terms of its satisfaction of the set of criteria X . In such a context the formula (14) may interpreted as follows:

- find a subset of criteria of the highest possible importance (expressed by μ) such that at the same time minimal satisfaction degree of all these criteria by the option F is as high as possible (15) (expressed by α)
- and take the minimum of these two degrees as the overall evaluation of the option F .

Now we will explain how various linguistic summaries discussed in the previous section may be interpreted using the Sugeno integral. The linguistic quantifier Q is still defined as in Zadeh's calculus as a fuzzy set in $[0,1]$, exemplified by (18). We will assume that Q is a regular monotone and nondecreasing quantifier:

$$\mu(0) = 0, \quad \mu(1) = 1 \quad (16)$$

$$x_1 \leq x_2 \Rightarrow \mu_Q(x_1) \leq \mu_Q(x_2) \quad (17)$$

exemplified by

$$\mu_Q(x) = \begin{cases} 1 & \text{for } x \geq 0.8 \\ 2x - 0.6 & \text{for } 0.3 < x < 0.8 \\ 0 & \text{for } x \leq 0.3 \end{cases} \quad (18)$$

The truth value of particular summaries is computed using the Sugeno integral (14). For simple types of summaries we are in a position to provide the interpretation similar to this given above for the MCDM. For this purpose we will identify the set of criteria X with a set of trends while an option F will be the whole time series under consideration characterized in terms of how well its trends satisfy P .

In what follows $|A|$ denotes the cardinality of set A , summarizers P and qualifiers R are identified with fuzzy sets modelling the linguistic terms they contain, X is the set

of all trends extracted from time series and $\text{time}(x_i)$ denotes duration of the trend x_i .

a) *Simple frequency based summaries defined by (8):*

The truth value of this type of summary may be expressed as $S_\mu(P)$ where

$$\mu(P_\alpha) = \mu_Q \left(\frac{|P_\alpha|}{|X|} \right) \quad (19)$$

Thus, referring to (15), the truth value is determined by looking for a subset of trends of the cardinality high enough as required by the semantics of the quantifier Q and such that all these trends “are P” to the highest possible degree.

b) *Extended frequency based summaries defined by (9):*

The truth value of this type of summary may be expressed as $S_\mu(P)$ where

$$\mu(P_\alpha) = \mu_Q \left(\frac{|(P \cap R)_\alpha|}{|R_\alpha|} \right) \quad (20)$$

c) *Simple duration based summaries defined with (10):*

The truth value of this type of summary may be expressed as $S_\mu(P)$ where

$$\mu(P_\alpha) = \mu_Q \left(\frac{\sum_{i:x_i \in P_\alpha} \text{time}(x_i)}{\sum_{i:x_i \in X} \text{time}(x_i)} \right) \quad (21)$$

Thus, referring to (15) the truth value is determined by looking for a subset of trends such that their total duration with respect to the duration of the whole time series is long enough as required by the semantics of the quantifier Q and such that all these trends “are P” to the highest possible degree.

d) *Extended duration based summaries defined with (11):*

The truth of this type of summary may be expressed as $S_\mu(P)$ where

$$\mu(P_\alpha) = \mu_Q \left(\frac{\sum_{i:x_i \in (P \cap R)_\alpha} \text{time}(x_i)}{\sum_{i:x_i \in R_\alpha} \text{time}(x_i)} \right) \quad (22)$$

Due to the properties (16)-(17) of the quantifiers employed it is obvious that all μ 's defined above for particular types of summaries satisfy the axioms (12) of the fuzzy measure.

V. EXAMPLE

Let us assume that from some given data we have extracted trends listed in Table I, e.g. using the algorithm shown in Figure 2. We assume the granulation of dynamics of change presented in Section II-A.

We can consider the following simple frequency based trend summary:

$$\text{Most of trends are decreasing} \quad (23)$$

In this summary *most* is the linguistic quantifier Q . The membership function is as in (18).

“Trends are decreasing” is a summarizer P with the membership function of the “decreasing” term given as in (24). Let us recall, that for brevity we identify summarizers and qualifiers with the linguistic terms they contain.

TABLE I
TRENDS EXTRACTED

id	dynamics of change (α in degrees)	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

$$\mu_P(\alpha) = \begin{cases} 0 & \text{for } \alpha \leq -65 \\ 0,066\alpha + 4.333 & \text{for } -65 < \alpha < -50 \\ 1 & \text{for } -50 \leq \alpha \leq -40 \\ -0.01\alpha - 1 & \text{for } -40 < \alpha < -20 \\ 0 & \text{for } \alpha \geq -20 \end{cases} \quad (24)$$

n is the number of all trends, i.e., in this example $n = |X|=9$.

The truth value of (23) is computed according to (14) and (19) that yields:

$$\begin{aligned} T(\text{Most of the trends are decreasing}) &= \\ &= \max_{\alpha_i \in \{a_1, \dots, a_n\}} \left(\alpha_i \wedge \mu_Q \left(\frac{|P_\alpha|}{|X|} \right) \right) = 0.511 \end{aligned}$$

If we assume the extended form, we may have the following summary:

$$\text{Most of short trends are decreasing} \quad (25)$$

Again, *most* is the linguistic quantifier Q with its membership function given as (18). “Trends are decreasing” is a summarizer P as in the previous example. “Trend is short” is the qualifier R . We define the membership function $\mu_R(t)$ as follows:

$$\mu_R(t) = \begin{cases} 1 & \text{for } t \leq 1 \\ -\frac{1}{2}t + \frac{3}{2} & \text{for } 1 < t < 3 \\ 0 & \text{for } t \geq 3 \end{cases} \quad (26)$$

The truth value of (25) is computed using the formula (14) and (20):

$$\begin{aligned} T(\text{Most of short trends are decreasing}) &= \\ &= \max_{\alpha_i \in \{a_1, \dots, a_n\}} \left(\alpha_i \wedge \mu_Q \left(\frac{|(P \cap R)_\alpha|}{|R_\alpha|} \right) \right) = 0.9 \end{aligned}$$

On the other hand, we may have the following simple duration based linguistic summary:

$$\text{Trends that took most time are slowly increasing} \quad (27)$$

“Trends are slowly increasing” is the summarizer P with the membership function $\mu_P(\alpha)$ defined as follows:

REFERENCES

- [1] I. Batyrshin (2002). On granular Derivatives and the solution of a Granular Initial Value Problem. In *International Journal Applied Mathematics and Computer Science*, 12(3):403-410.
- [2] I. Batyrshin, L. Sheremetov. Perception Based Functions in Qualitative Forecasting. (to appear)
- [3] P. Bosc P., L. Lietard, O. Pivert (2003). Sugeno fuzzy integral as a basis for the interpretation of flexible queries involving monotonic aggregates. In *Information Processing and Management*, 39(2):287–306.
- [4] M. Grabisch (1998). Fuzzy integral as a flexible and interpretable tool of aggregation. In Bouchon-Meunier B. (ed.) *Aggregation and Fusion of Imperfect Information*, Studies in Fuzziness and Soft Computing, Heidelberg, New York: Physica–Verlag, 51–72.
- [5] J. Colomer, J. Melendez, J. L. de la Rosa, J. Augilar (1997). A qualitative/quantitative representation of signals for supervision of continuous systems. In *Proceedings of the European Control Conference -ECC97*, Brussels.
- [6] B. Hugeney, B. Bouchon-Meunier (2001). Time-Series Segmentation and Symbolic Representation, from Process-Monitoring to Data-Mining. In *LNCS 2206*:118-123.
- [7] J. Kacprzyk, A. Wilbik, S. Zadrożny (2006). Linguistic summarization of trends: an approach. (in press).
- [8] J. Kacprzyk and R.R. Yager (2001). Linguistic summaries of data using fuzzy logic. In *International Journal of General Systems*, 30:33-154.
- [9] J. Kacprzyk, R.R. Yager and S. Zadrożny (2000). A fuzzy logic based approach to linguistic summaries of databases. In *International Journal of Applied Mathematics and Computer Science*, 10:813-834.
- [10] J. Kacprzyk, S. Zadrożny (2005). Linguistic database summaries and their protoforms: toward natural language based knowledge discovery tools. In *Information Sciences* 173:281-304.
- [11] J. Kacprzyk, S. Zadrożny (2005). Fuzzy linguistic data summaries as a human consistent, user adaptable solution to data mining. In B. Gabrys, K. Leiviska, J. Strackeljan (Eds.) *Do Smart Adaptive Systems Exist?* Springer, Berlin Heidelberg New York, 321-339.
- [12] J. Sklansky and V. Gonzalez (1980) Fast polygonal approximation of digitized curves. In *Pattern Recognition* 12(5):327-331.
- [13] R.R. Yager (1982). A new approach to the summarization of data. In *Information Sciences*, 28:69-86.
- [14] L.A. Zadeh (1983). A computational approach to fuzzy quantifiers in natural languages. In *Computers and Mathematics with Applications*, 9, 149-184.
- [15] L.A. Zadeh (2002). A prototype-centered approach to adding deduction capabilities to search engines – the concept of a protoform. BISC Seminar, University of California, Berkeley.

$$\mu_P(\alpha) = \begin{cases} 0 & \text{for } \alpha \leq 5 \\ 0.1\alpha - 0.5 & \text{for } 5 < \alpha < 15 \\ 1 & \text{for } 15 \leq \alpha \leq 20 \\ -0.05\alpha + 2 & \text{for } 20 < \alpha < 40 \\ 0 & \text{for } \alpha \geq 40 \end{cases} \quad (28)$$

The linguistic quantifier *most* is defined as previously. The truth value of (27) is computed via the formula (14) and (21) and we obtain:

$$\begin{aligned} &T(\text{Trends that took } \textit{most} \text{ time are } \textit{slowly increasing}) \\ &= \max_{\alpha_i \in \{a_1, \dots, a_n\}} \left(\alpha_i \wedge \mu_Q \left(\frac{\sum_{x_i \in P_\alpha} \text{time}(x_i)}{\sum_{i: x_i \in X} \text{time}(x_i)} \right) \right) \\ &= 0.733 \end{aligned}$$

Finally, we may consider an extended form of duration based summaries, here exemplified by:

$$\text{Trends with a low variability that took most of the time are } \textit{slowly increasing} \quad (29)$$

Again, *most* is the linguistic quantifier and “*trends are slowly increasing*” is summarizer *P*, with a membership function defined as in the previous example. “*Trends have a low variability*” is the qualifier *R*. The membership function $\mu_R(v)$ is given as follows:

$$\mu_R(v) = \begin{cases} 1 & \text{for } v \leq 0.2 \\ -5v + 2 & \text{for } 0.2 < v < 0.4 \\ 0 & \text{for } v \geq 0.4 \end{cases} \quad (30)$$

The truth value of (29) is computed according to the formula (14) and (22) and we obtain:

$$\begin{aligned} &T(\text{Trends with low variability that took most of the time are } \textit{slowly increasing}) \\ &= \max_{\alpha_i \in \{a_1, \dots, a_n\}} \left(\alpha_i \wedge \mu_Q \left(\frac{\sum_{i: x_i \in (P \cap R)_\alpha} \text{time}(x_i)}{\sum_{i: x_i \in R_\alpha} \text{time}(x_i)} \right) \right) \\ &= 0.75 \end{aligned}$$

VI. CONCLUDING REMARKS

We have proposed to adapt the linguistic summarization of databases to the case of time series. The basic idea consists in identifying in time series trends that are characterized by a set of attributes. Then such a set of trends is directly amenable to the linguistic summarization. The specificity of time series calls for a new type of summaries that cannot be easily cast in the original framework of linguistic summaries as proposed by Yager. Due to that we propose to use the Sugeno integral to model quantifier based aggregation for all types of summaries. The basic idea is very similar and inspired by the work of Bosc et al. [3].