

On the evaluation of the linguistic summarization of temporally focused time series using a measure of informativeness

Anna Wilbik and Janusz Kacprzyk

Systems Research Institute

Polish Academy of Sciences

ul. Newelska 6, 01-447 Warsaw, Poland

Email: wilbik@ibspan.waw.pl, kacprzyk@ibspan.waw.pl

Abstract—We extend our previous works of deriving linguistic summaries of time series using a fuzzy logic approach to linguistic summarization. We proceed towards a multicriteria analysis of summaries by assuming as a quality criterion Yager’s measure of informativeness of classic and temporal protoforms that combines in a natural way the measures of truth, focus and specificity, to obtain a more advanced evaluation of summaries. The use of the informativeness measure for the purpose of a multicriteria evaluation of linguistic summaries of time series seems to be an effective and efficient approach, yet simple enough for practical applications. Results on the summarization of quotations of an investment (mutual) fund are very encouraging.

I. INTRODUCTION

WITH this paper we continue our previous works (cf. Kacprzyk, Wilbik, Zadrozny [1], [2], [3] or Kacprzyk, Wilbik [4], [5], [6]) which deal with the problem of how to effectively and efficiently support humans in making decisions concerning investments in some financial, notably in investment (mutual) funds. Decision makers are here mainly interested in future gains/losses. However, we follow the decision support paradigm, that is, we assume that the users are autonomous and we only support, not replace, him/her. We do not intend to forecast the future daily prices.

The available information concerns the history, or past, and this implies some problems. Basically in all investment decisions the future is the most important, and the past is irrelevant. But, we know only the the past, and the future is completely unknown. Behavior of the human being is to a large extent driven by his/her (already known) past experience. People usually tend to assume that what has happened in the past will also happen (to some, maybe large extent) in the future. By the way, this is the underlying assumption behind the statistical methods too! That attitude clearly implies that the past can be employed to help the human decision maker find a good solution. We follow here this path, i.e. we present a method to subsume the past, to be more specific the past performance of an investment (mutual) fund, by presenting results in a very human consistent way, using natural language statements.

This line of reasoning has often been articulated by many well known investment practitioners, and one can quote here

some more relevant opinions. In any information leaflets of investment funds, one may always notice a disclaimer stating that “Past performance is no indication of future returns” which is true. However, on the other hand, for instance, in a well known posting “Past Performance Does Not Predict Future Performance” [7], they state something that may look strange in this context, namely: “...according to an Investment Company Institute study, about 75% of all mutual fund investors mistakenly use short-term past performance as their primary reason for buying a specific fund”. But, in an equally well known posting “Past performance is not everything” [8], they state: “...disclaimers apart, as a practice investors continue to make investments based on a schemes past performance. To make matters worse, fund houses are only too pleased to toe the line by actively advertising the past performance of their schemes leading investors to conclude that it is the single-most important parameter (if not the most important one) to be considered while investing in a mutual fund scheme”.

As strange as this may be, we may ask ourselves why it is so. Again, in a well known posting “New Year’s Eve: Past performance is no indication of future return” [9], they say “...if there is no correlation between past performance and future return, why are we so drawn to looking at charts and looking at past performance? I believe it is because it is in our nature as human beings ...because we don’t know what the future holds, we look toward the past ...”.

And, continuing along this line of reasoning, we can find many other examples of similar statements supporting our position. For instance, Myers [10] says: “...Does this mean you should ignore past performance data in selecting a mutual fund? No. But it does mean that you should be wary of how you use that information ...*Lousy performance in the past is indicative of lousy performance in the future...*”. And, further: Bogle [11] states: “... there is an important role that past performance can play in helping you to make your fund selections. While you should disregard a single aggregate number showing a fund’s past long-term return, you can learn a great deal by studying the *nature of its past returns*. Above all, look for consistency.”. In [12], we find: “While

past performance does not necessarily predict future returns, it can tell you how volatile a fund has been". In the popular "A 10-step guide to evaluating mutual funds" [13], they say in the last, tenth, advise: "Evaluate the funds performance. Every fund is benchmarked against an index like the S&P500, BSE 200, etc. Investors should compare fund performance over varying time frames vis-a-vis both the benchmark index and peers. Carefully evaluate the funds performance across market cycles particularly the downturns".

We can quote more, and basically all of them stress the importance of looking at the past to help make future decisions, and also generally advocate a more comprehensive look not focused on single values but a very essence of past behavior and returns.

We have followed this line of reasoning in our past papers (cf. Kacprzyk, Wilbik, Zadrozny [1], [2], [3] or Kacprzyk, Wilbik [4], [5], [6]), i.e. to try to find a human consistent, fuzzy quantifier based scheme for a linguistic summarization of the past in terms of various aspects of how the time series representing daily quotations of the investment fund(s) behave. However, we have mainly concentrated on a sheer absolute performance, i.e. the time evolution of the quotations themselves. This may be relevant, and sometimes attractive to the users who can see a summary of their gains/loses and their temporal evolution. One can also use a maybe more realistic approach to take into account benchmarks of the particular funds as points of departure which does not change the essence.

Though the use of linguistic data summaries of past performance of the time series representing mutual fund quotations does take into account the importance (or "value") of time, in this paper we will go deeper into this issue by using some results from psychology, cognitive sciences and human decision making. Basically, we will employ some results by Ariely and Zakay [14] who consider the role of time in decision making.

In our case, those psychological analyses served the purpose of suggesting, and/or justifying a new types of protoforms of linguistic summaries of time series. Basically, in most of our works (cf. Kacprzyk, Wilbik, Zadrozny [1], [2], [3] or Kacprzyk, Wilbik [4], [5], [6]) we have used the following protoforms of the linguistic summaries of times series: "Among all y 's, Q are P ", exemplified by "among all segments (of the time series) most are slowly increasing", and "Among all R segments, Q are P ", exemplified by "among all short segments almost all are quickly decreasing". Notice that we took into account, and so to say assign the same weights, the entire time series, i.e. all the segment.

Since in our case the analysis of time series is a highly human focused activity because its very purpose is to provide a human decision maker with some support for making (future) decision, we should take into account some inherent characteristics of time series and their evaluations that are consistent with the human perception of their relevance for the decision making process. One of the crucial aspects in this respect, which will be considered here is the importance

of time in the sense that means and ends, like decisions and outcomes, have a carrying relevance and impact depending on the time moment when they occur. Basically, in virtually all cases what occurs in a more immediate past is more relevant and meaningful than what has occurred earlier. This temporal relationships change both the decisions and their evaluation as has been shown in psychology (cf. Ariely and Zakay [14] or Rachlin [15]). Among many approaches one can mention, for instance, a so called *temporal construal theory* by Liberman and Trope [16] who have shown that options are evaluated differently depending on time instants they come into question. They introduce the two main characteristics of options: desirability, which refers to long time wishes or intentions that are far away of their implementation of a decision option, and feasibility, which refers to a short term, close to the implementation characteristics. One can mention other works concerned with similar issues. It should be noted that this fact has already been reflected in (dynamic, or multistage) decision making and control models in which discounting is widely used.

In our context, we proposed (cf. Kacprzyk, Wilbik [17]) to take into account some of those psychological findings related to the importance of time by using different protoforms of linguistic summaries of times series, called *temporal linguistic summaries*. We consider two types of temporal protoforms: " E_T among all y 's Q are P ", exemplified by "Recently, among all segments, most are slowly increasing", and " E_T among all Ry 's Q are P ", exemplified by "Initially, among all short segments, most are quickly decreasing"; they both go beyond the classic Zadeh's protoforms.

The analysis of time series data involves different elements but we concentrate on the specifics of our approach. First, we have to identify the consecutive parts of time series within which the data exhibit some uniformity as to their variability. Some variability must here be neglected, under an assumed granularity. Here, these consecutive parts of a time series are called trends (or segments), and described by straight line segments. That is, we perform first a piece-wise linear approximation of a time series and present time series data as a sequence of trends. The (linguistic) summaries of time series refer to the (linguistic) summaries of (partial) trends as meant above. For the construction of a piecewise linear approximation, we use a modified version of the Sklansky and Gonzalez algorithm (cf. [18]) though many other methods can be used cf. Keogh et al. [19], [20].

The next step is an aggregation of the (characteristic features of) consecutive trends over an entire time span (horizon) assumed. We follow the idea initiated by Yager [21], [22] and then shown more profoundly and in an implementable way in Kacprzyk and Yager [23], and Kacprzyk, Yager and Zadrozny [24], [25], that the most comprehensive and meaningful will be a linguistic quantifier driven aggregation resulting in linguistic summaries of classic protoform exemplified by "Most trends are short" or "Most long trends are increasing" and temporal protoform "Recently most trends are increasing" or "Recently most short trends are increasing".

These summaries are easily derived and interpreted using Zadehs fuzzy logic based calculus of linguistically quantified propositions. A new quality, and an increased generality was obtained by using Zadehs [26] protoforms as proposed by Kacprzyk and Zadrozny [27].

Here we employ the classic Zadehs fuzzy logic based calculus of linguistically quantified propositions in which the degree of truth (validity) is the most obvious and important quality criterion. Some other quality criteria like a degree of specificity, focus, fuzziness, etc. have also been proposed by Kacprzyk and Wilbik [28], [6], [5], [29]. The results obtain clearly indicate that multiple quality criteria of linguistic summaries of time series should be taken into account, and this makes the analysis obviously much more difficult.

As the first step towards an intended comprehensive multi-criteria assessment of linguistic summaries of time series, we propose here a very simple, effective and efficient approach, namely to use quite an old, maybe classic Yagers [30] proposal on an informativeness measure of a linguistic summary which combines, via an appropriate aggregation operator, the degree of truth, focus and specificity.

We illustrate our analysis on a linguistic summarization of daily quotations over an 8 year period of an investment (mutual) fund. We present the characteristic features of trends derived under some reasonable granulations, variability, trend duration, etc.

The paper is in line with some other modern approaches to linguistic summarization of time series. First, one should refer to the *SumTime* project coordinated by the University of Aberdeen, an EPSRC Funded Project for Generating Summaries of Time Series Data¹ in which English summary descriptions of a time series data set are sought by using advanced time series and NLG (natural language generation) technologies [31]. However, the linguistic descriptions obtained do not reflect an inherent imprecision (fuzziness) as in our approach. A relation between linguistic data summaries and NLG is discussed by Kacprzyk and Zadrozny [32], [33].

II. LINGUISTIC DATA SUMMARIES

As a *linguistic summary of data (base)* we understand a (usually short) sentence (or a few sentences) that captures the very essence of the set of data, that is numeric, large, and because of its size is not comprehensible for human being.

We use Yager's basic approach [21]. A linguistic summary includes: (1) a summarizer P (e.g. *low* for attribute *salary*), (2) a quantity in agreement Q , i.e. a linguistic quantifier (e.g. *most*), (3) truth (validity) \mathcal{T} of the summary and optionally, (4) a qualifier R (e.g. *young* for attribute *age*).

Thus, a linguistic summary may be exemplified by

$$\mathcal{T}(\text{most of employees earn low salary}) = 0.7 \quad (1)$$

or in richer (extended) form, including a qualifier (e.g. *young*), by

$$\mathcal{T}(\text{most of young employees earn low salary}) = 0.82 \quad (2)$$

¹cf. www.csd.abdn.ac.uk/research/sumtime/

Thus, basically the core of a linguistic summary is a linguistically quantified proposition in the sense of Zadeh [26] which may be written, respectively as

$$Qy's \text{ are } P \qquad QRy's \text{ are } P \quad (3)$$

III. LINGUISTIC SUMMARIES OF TRENDS

In our first approach we summarize the trends (segments) extracted from time series. Therefore as the first step we need to extract the segments. We assume that segment is represented by a fragment of straight line, because such segments are easy for interpretation.

There are many algorithms for the piecewise linear segmentation of time series data, including e.g. on-line (sliding window) algorithms, bottom-up or top-down strategies (cf. Keogh [19], [20]).

We consider the following three features of (global) trends in time series: (1) dynamics of change, (2) duration, and (3) variability. By *dynamics of change* we understand the speed of change of the consecutive values of time series. It may be described by the slope of a line representing the trend, represented by a linguistic variable. *Duration* is the length of a single trend, and is also represented by a linguistic variable. *Variability* describes how "spread out" a group of data is. We compute it as a weighted average of values taken by some measures used in statistics: (1) the range, (2) the interquartile range (IQR), (3) the variance, (4) the standard deviation, and (5) the mean absolute deviation (MAD). This is also treated as a linguistic variable.

For practical reasons for all we use a fuzzy granulation (cf. Bathyrshin at al. [34], [35]) to represent the values by a small set of linguistic labels as, e.g.: increasing, slowly increasing, constant, slowly decreasing, decreasing. These values are equated with fuzzy sets.

For clarity and convenience we employ Zadeh's [36] protoforms for dealing with linguistic summaries [27]. A protoform is defined as a more or less abstract prototype (template) of a linguistically quantified proposition. We have two types of protoforms of linguistic summaries of trends:

– a short form:

$$\text{Among all segments, } Q \text{ are } P \quad (4)$$

e.g.: "Among all segments, *most* are *slowly increasing*".

– an extended form:

$$\text{Among all } R \text{ segments, } Q \text{ are } P \quad (5)$$

e.g.: "Among all *short* segments, *most* are *slowly increasing*".

We can extend our protoforms given in (4) and (5) by adding a temporal expression, E_T , like: "recently", "in the very beginning" or "in May 2010", "initially", etc. (cf. Kacprzyk, Wilbik [17]). The temporal protoforms can have the following forms:

- a simple (short) protoform:

$$E_T \text{ among all segments, } Q \text{ are } P \quad (6)$$

e.g.: “Recently, among all segments, *most* are *slowly increasing*”.

- an extended protoform:

$$E_T \text{ among all } R \text{ segments, } Q \text{ are } P \quad (7)$$

e.g.: “Initially, among all *short* segments, *most* are *slowly increasing*”.

The quality of linguistic summaries can be evaluated in many different ways, eg. using the degree of truth, specificity, appropriateness or others.

Yager [30] proposed measure of informativeness, a measure that evaluates the amount of information hidden in the summary. This measure is interesting as it aggregates some of previously mentioned quality criteria, namely the truth value, degree of specificity and degree of focus in the case of extended form summaries. Now we will present shortly those 3 measures.

A. Truth Value

The truth value (a degree of truth or validity), introduced by Yager in [21], is the basic criterion describing the degree of truth (from $[0, 1]$) to which a linguistically quantified proposition equated with a linguistic summary is true.

Using Zadeh’s calculus of linguistically quantified propositions [26] it is calculated in dynamic context using the same formulas as in the static case. Thus, the truth value is calculated for the simple and extended form as, respectively:

$$\mathcal{T}(\text{Among all } y\text{'s, } Q \text{ are } P) = \mu_Q \left(\frac{1}{n} \sum_{i=1}^n \mu_P(y_i) \right) \quad (8)$$

$$\begin{aligned} \mathcal{T}(\text{Among all } Ry\text{'s, } Q \text{ are } P) &= \\ &= \mu_Q \left(\frac{\sum_{i=1}^n \mu_R(y_i) \wedge \mu_P(y_i)}{\sum_{i=1}^n \mu_R(y_i)} \right) \end{aligned} \quad (9)$$

where μ_P , μ_R and μ_Q are membership functions of fuzzy set representing summarizer, qualifier and linguistic quantifier, respectively. \wedge is the minimum operation (more generally it can be another appropriate operator, notably a t -norm). In Kacprzyk, Wilbik and Zadrozny [37] results obtained by using different t -norms were compared. Various t -norms can be in principle used in Zadeh’s calculus but clearly their use may result in different results of the linguistic quantifier driven aggregation. It seems that the minimum operation is a good choice since it can be easily interpreted and the numerical values correspond to the intuition.

The computation of truth values of temporal summaries is very similar to the previous case. We only need to consider a temporal expression as an additional external qualifier, as the temporal expression limits the universe of interest to those trends (segments) only that occur on the time axis described by a fuzzy set modeling the expression E_T . We compute the proportion of segments in which “trend is P” and occurs in E_T to those that occur in E_T . Next, we compute the degree to which this proportion is Q .

The truth value of the simple temporal protoform (6) is computed as:

$$\begin{aligned} \mathcal{T}(E_T \text{ among all } y\text{'s, } Q \text{ are } P) &= \\ &= \mu_Q \left(\frac{\sum_{i=1}^n \mu_{E_T}(y_i) \wedge \mu_P(y_i)}{\sum_{i=1}^n \mu_{E_T}(y_i)} \right) \end{aligned} \quad (10)$$

where $\mu_{E_T}(y_i)$ is the degree to which a trend (segment) occurs during the time span described by E_T .

Similarly we compute the truth of the extended temporal protoform (7) as:

$$\begin{aligned} \mathcal{T}(E_T \text{ Among all } Ry\text{'s, } Q \text{ are } P) &= \\ &= \mu_Q \left(\frac{\sum_{i=1}^n \mu_{E_T}(y_i) \wedge \mu_R(y_i) \wedge \mu_P(y_i)}{\sum_{i=1}^n \mu_{E_T}(y_i) \wedge \mu_R(y_i)} \right) \end{aligned} \quad (11)$$

A natural question emerges of how to compute $\mu_{E_T}(y_i)$. Let $\mu_{E_T}(t)$ be a membership function of a fuzzy set representing a linguistic variable E_T . We assume that the time span considered is normalized, i.e. $t \in [0, 1]$, the first observation is made for $t = 0$ and the last for $t = 1$. Let us consider a segment y_i , starting at time a and terminating at time b , $0 \leq a < b \leq 1$. Then

$$\mu_{E_T}(y_i) = \frac{1}{b-a} \int_a^b \mu_{E_T}(t) dt \quad (12)$$

and we can interpret this value as the average membership degree of E_T in $[a, b]$. Graphically it can be represented as the gray striped area divided by the striped area in Figure 1.

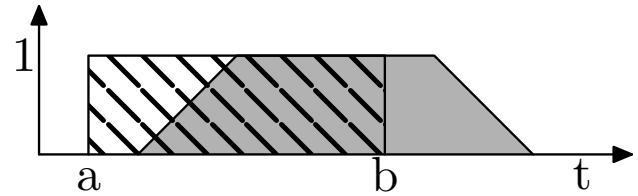


Fig. 1. Graphical presentation of $\mu_{E_T}(y_i)$

B. Degree of Specificity

The concept of specificity provides a measure of the amount of information contained in a fuzzy subset or possibility distribution. The specificity measure evaluates the degree to which a fuzzy subset points to one and only one element as its member [38].

We will consider the original Yagers proposal [38], in which specificity measures the degree to which a fuzzy subset contains one and only one element. The measure of specificity is a measure $Sp: I^X \rightarrow I$, $I \in [0, 1]$ if it has the following properties: (1) $Sp(A) = 1$ if and only if $A = \{x\}$, (is a singleton set), (2) $Sp(\emptyset) = 0$, and (3) $\frac{\partial Sp(A)}{\partial \alpha_1} > 0$ and $\frac{\partial Sp(A)}{\partial \alpha_j} \leq 0$ for all $j \geq 2$.

In [39] Yager proposed a measure of specificity as

$$Sp(A) = \int_0^{\alpha_{max}} \frac{1}{card(A_\alpha)} d\alpha \quad (13)$$

where α_{max} is the largest membership grade in A , A_α is the α -level set of A , (i.e. $A_\alpha = \{x : \mu_A(x) \geq \alpha\}$) and $card(A_\alpha)$ is the number of elements in A_α .

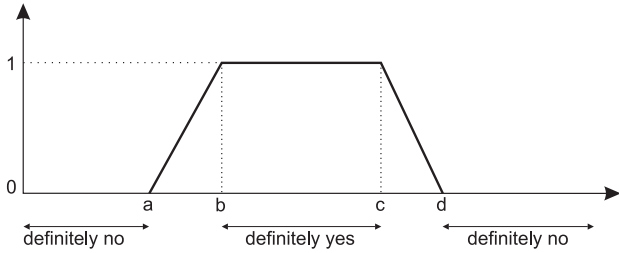


Fig. 2. A trapezoidal membership function of a set

In our summaries to define the membership functions of the linguistic values we use trapezoidal functions, as they are sufficient in most applications [40]. Moreover, they can be very easily interpreted and defined by a user not familiar with fuzzy sets and logic, as in Figure 2. To represent a fuzzy set with a trapezoidal membership function we need to store only four numbers, a , b , c and d . Usage such a definition of a fuzzy set is a compromise between cointension and computational complexity. In such a case measure of specificity of a fuzzy set A

$$Sp(A) = 1 - \frac{c + d - (a + b)}{2} \quad (14)$$

C. Degree of Focus

The very purpose of a degree of focus is to limit the search for the best linguistic summaries by taking into account some additional information in addition to truth values. The extended protoform linguistic summaries (5) does limit by itself the search space as the search is performed in a limited subspace of all (most) trends that fulfill an additional condition specified by qualifier R . The very essence of the degree of focus introduced in this paper is to give the proportion of trends satisfying property R to all trends extracted from the time series. It provides a measure that, in addition to the basic truth value, can help control the process of discarding non-promising linguistic summaries.

The *degree of focus* is similar in spirit to a degree of covering, described above, but it measures how many trends fulfill property R . The degree of focus makes obviously sense for the extended protoform summaries only, and is calculated as (cf. Kacprzyk and Wilbik [29]):

$$d_{foc}(\text{Among all } Ry\text{'s, } Q \text{ are } P) = \frac{1}{n} \sum_{i=1}^n \mu_R(y_i) \quad (15)$$

In our context, the degree of focus describes how many trends extracted from a given time series fulfill qualifier R in comparison to all extracted trends. If the degree of focus is high, then we can be sure that such a summary concerns many trends, so that it is more general. However, if the degree of focus is low, we may be sure that such a summary describes a (local) pattern seldom occurring.

The formula for the degree of focus for the extended temporal protoform requires small changes. The temporal expression may be treated as the external qualifier, and we can compute the proportion of trends satisfying property R in the E_T time span to all trends occurring in that time span. So the degree of focus of extended temporal protoform summaries (7) is computed as:

$$\begin{aligned} d_{foc}(E_T \text{ among all } Ry\text{'s, } Q \text{ are } P) &= \\ &= \frac{\sum_{i=1}^n \mu_{E_T}(y_i) \wedge \mu_R(y_i)}{\sum_{i=1}^n \mu_{E_T}(y_i)} \end{aligned} \quad (16)$$

Here also the degree of focus help us distinguish more general summaries from those describing a (local) pattern seldom occurring.

As we wish to discover a more general, global relationship, we can eliminate the linguistic summaries that concern a small number of trends only. The degree of focus may be used to eliminate the whole groups of extended form summaries for which qualifier R limits the set of possible trends to, for instance, 5%. Such summaries, although they may be very true, will not be representative.

We could think also about an additional measure similar to the degree of focus for the temporal protoforms – a degree of focus of temporal expression. This degree could measure how many trends extracted from a given time series occurs in the time span described by E_T in comparison to all extracted trends. Hence, for the simple and extended temporal protoform summaries we have:

$$d_{E_T}(E_T \text{ among all } Ry\text{'s, } Q \text{ are } P) = \frac{1}{n} \sum_{i=1}^n \mu_{E_T}(y_i) \quad (17)$$

D. Measure of Informativeness

The idea of a measure of informativeness (cf. Yager, Ford and Canas [30]) may be summarized as follows. Suppose we have a data set, whose elements are from measurement space X . One can say that the data set itself is its own most informative description, and any other summary implies a loss of information. So, a natural question is whether a particular summary is informative, and to what extent.

Yager et. al [30] proposed the following *measure of informativeness* of a simple protoform summary

$$\begin{aligned} I(\text{Among all } y\text{'s } Q \text{ are } P) &= \\ &= (\mathcal{T} \cdot Sp(Q) \cdot Sp(P)) \vee ((1 - \mathcal{T}) \cdot Sp(Q^c) \cdot Sp(P^c)) \end{aligned} \quad (18)$$

where P^c is the negation of P , i.e. $\mu_{P^c}(\cdot) = 1 - \mu_P(\cdot)$ and Q^c is the negation of Q , i.e. $\mu_{Q^c}(\cdot) = 1 - \mu_Q(\cdot)$. $Sp(Q)$ is specificity of Q , similarly it is calculated for Q^c , P and P^c .

For the extended protoform summary we propose the following measure (cf. Kacprzyk and Wilbik [41]):

$$\begin{aligned} I(\text{Among all } Ry\text{'s } Q \text{ are } P) &= \\ &= (\mathcal{T} \cdot Sp(Q) \cdot Sp(P) \cdot Sp(R) \cdot d_{foc}) \\ &\quad \vee ((1 - \mathcal{T}) \cdot Sp(Q^c) \cdot Sp(P^c) \cdot Sp(R) \cdot d_{foc}) \end{aligned} \quad (19)$$

where d_{foc} is the degree of focus of the summary, $Sp(R)$ is specificity of qualifier R and the rest is defined as previously.

The measure of informativeness of the simple temporal protoform summary is calculated as:

$$\begin{aligned} I(E_T \text{ among all } y\text{'s } Q \text{ are } P) &= \\ &= (\mathcal{T} \cdot Sp(Q) \cdot Sp(P) \cdot Sp(E_T) \cdot d_{E_T}) \\ &\quad \vee ((1 - \mathcal{T}) \cdot Sp(Q^c) \cdot Sp(P^c) \cdot Sp(E_T) \cdot d_{E_T}) \end{aligned} \quad (20)$$

where $Sp(E_T)$ is the specificity of the temporal expression and d_{E_T} is the degree of focus of temporal expression defined as in Eq. (17).

The measure of informativeness of the extended temporal protoform summary is calculated as:

$$\begin{aligned} I(E_T \text{ among all } Ry\text{'s } Q \text{ are } P) &= \\ &= (\mathcal{T} \cdot Sp(Q) \cdot Sp(P) \cdot Sp(E_T) \cdot Sp(R) \cdot d_{foc} \cdot d_{E_T}) \\ &\quad \vee ((1 - \mathcal{T}) \cdot Sp(Q^c) \cdot Sp(P^c) \cdot Sp(E_T) \cdot Sp(R) \cdot \\ &\quad \cdot d_{foc} \cdot d_{E_T}) \end{aligned} \quad (21)$$

Here in those formulas different values are aggregated by the product. We could think of using instead of the product other t -norms. However, for example, the minimum would ignore all values that are smaller than the largest one, and the Łukasiewicz t -norm tends to be very small if we aggregate many numbers. Moreover, the product may be a natural choice taking into account many results from, for instance, decision analysis and mathematical economics.

IV. NUMERICAL RESULTS

The method proposed was tested on data on quotations of an investment (mutual) fund that invests at least 50% of assets in shares listed at the Warsaw Stock Exchange.

Data shown in Figure 3 were collected from January 2002 until the December 2009 with the value of one share equal to PLN 12.06 in the beginning of the period to PLN 35.82 at the end of the time span considered (PLN stands for the Polish Zloty). The minimal value recorded was PLN 9.35 while the maximal one during this period was PLN 57.85. The biggest daily increase was equal to PLN 2.32, while the biggest daily decrease was equal to PLN 3.46. We illustrate the method proposed by analyzing the absolute performance of a given investment fund, and not against benchmarks, for illustrativeness.

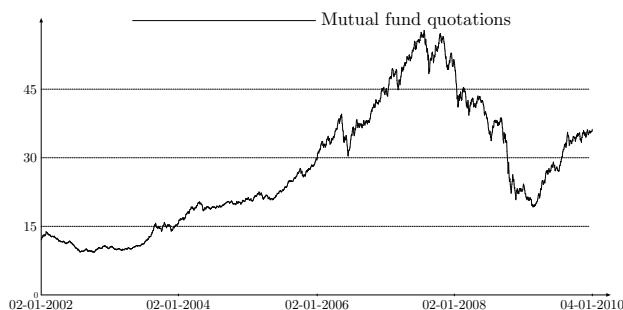


Fig. 3. Daily quotations of an investment fund in question

We obtain 362 extracted trends, with the shortest of 1 time unit only, and the longest – 71 time units. We assume 3 labels

only for each attribute: short, medium and long for duration, increasing, constant and decreasing for dynamics and low, moderate and high for variability. The use of linguistic values in the summaries is clearly a reflection of a natural information granulation.

In Table I there are presented the most valid summaries of the classic protoforms. They are ordered according to the degree of truth and then the degree of focus.

TABLE I
SUMMARIES OF THE CLASSIC PROTOFORM

linguistic summary	\mathcal{T}	d_{foc}	I
Among all y, most are short	1.0000	1.0000	0.4675
Among all moderate y, most are short	1.0000	0.3453	0.0969
Among all decreasing y, most are short	1.0000	0.2267	0.0604
Among all increasing y, most are short	1.0000	0.1688	0.0450
Among all medium y, most are constant	1.0000	0.1394	0.0429
Among all medium y, most are constant and high	1.0000	0.1394	0.0778
Among all medium y, most are high	1.0000	0.1394	0.0349
Among all medium and constant y, most are high	1.0000	0.1366	0.0794
Among all medium and high y, most are constant	1.0000	0.1222	0.0780
Among all high y, most are short	0.8453	0.5789	0.1601
Among all constant y, most are short	0.8341	0.6045	0.2027
Among all high y, most are constant	0.8164	0.5789	0.1565
Among all constant y, most are high	0.7564	0.6045	0.1514

We may notice that the first summary has a very big value of the measure of the informativeness, and this summary is of a simple protoform. It is very informative. Also the last four summaries presented in Table I are interesting, as their values of this measure are quite high. Those summaries do not have the truth value equal to 1, but nevertheless they are also true, moreover they have very high values of degree of focus, indicating that these summaries describe pattern which are quite often occurring.

In Table II we may see the temporal summaries describing that time series. We may notice the summaries of the first few years after the fund was established (initially), then the middle time (in the middle), and the last two summaries describe more or less time quotations from autumn 2007, when the financial crisis started.

The obtained summaries are divided into 2 groups, each describing separate period. The first group describes what was happening initially. First 4 summaries have high values of the measure of informativeness. Especially interesting is the summary “initially among all y, most are constant and very high”, it has the value of this measure higher than a summary with only one of the linguistic values used in the summary.

In the second group, only one summary stands out – “middle among all y, most are constant”, with a value of the measure of informativeness over twice as big than the other values. We may notice that this value is also the biggest for all summaries presented in Table II. It is partially so, because the temporal expression “in the middle” describes the longest period.

In the last group describing the last 2 years – the time of financial crisis, we obtained just 2 summaries, from which the

TABLE II
SUMMARIES OF THE TEMPORAL PROTOFORM

linguistic summary	\mathcal{T}	d_{foc}	I
initially among all constant y, most are very high	1.0000	1.0000	0.0329
initially among all y, most are constant	1.0000	1.0000	0.0387
initially among all y, most are constant and very high	1.0000	1.0000	0.0770
initially among all y, most are very high	1.0000	1.0000	0.0383
initially among all very high y, most are constant	1.0000	0.8089	0.0266
initially among all long and constant y, most are very high	1.0000	0.3566	0.0213
initially among all long y, most are constant	1.0000	0.3566	0.0097
initially among all long y, most are constant and very high	1.0000	0.3566	0.0192
initially among all long y, most are very high	1.0000	0.3566	0.0096
initially among all medium and constant y, most are very high	1.0000	0.3292	0.0214
initially among all medium and very high y, most are constant	1.0000	0.3292	0.0215
initially among all medium y, most are constant	1.0000	0.3292	0.0107
initially among all medium y, most are constant and very high	1.0000	0.3292	0.0212
initially among all medium y, most are very high	1.0000	0.3292	0.0106
initially among all very long y, most are constant	1.0000	0.3230	0.0069
initially among all long and very high y, most are constant	1.0000	0.2947	0.0177
initially among all very long and very high y, most are constant	1.0000	0.1939	0.0105
initially among all high y, most are constant	1.0000	0.1911	0.0059
initially among all very long and high y, most are constant	1.0000	0.1291	0.0068
initially among all high y, most are very long	0.7518	0.1911	0.0028
initially among all high y, most are very long and constant	0.7518	0.1911	0.0073
middle among all very high y, most are constant	1.0000	0.4860	0.0481
middle among all medium y, most are constant	1.0000	0.3205	0.0312
middle among all high y, most are short	1.0000	0.2436	0.0249
middle among all medium and very high y, most are constant	1.0000	0.2422	0.0476
middle among all short and very high y, most are constant	1.0000	0.1094	0.0228
middle among all y, most are constant	0.9659	1.0	0.1124
middle among all medium y, most are very high	0.9113	0.3205	0.0281
middle among all short y, most are constant	0.8506	0.4794	0.0447
middle among all medium and constant y, most are very high	0.8484	0.2840	0.0470
middle among all moderate y, most are short	0.8188	0.2741	0.0200
middle among all short and moderate y, most are constant	0.8084	0.1945	0.0301
middle among all moderate y, most are constant	0.8010	0.2741	0.0179
from the crisis begin among all medium y, most are constant	1.0000	0.2273	0.0513
from the crisis begin among all decreasing y, most are very short	0.9887	0.1438	0.0294

first one is more informative than the other, nevertheless both seems to be interesting for the experts.

V. CONCLUDING REMARKS

We extended our approach to the linguistic summarization of time series towards a multicriteria analysis of classic and temporal summaries by assuming as a quality criterion Yager's measure of informativeness that combines in a natural way the measures of truth, focus and specificity. Results on the summarization of quotations of an investment (mutual) fund are very encouraging.

REFERENCES

- [1] J. Kacprzyk, A. Wilbik, and S. Zadrozny, "Linguistic summarization of trends: a fuzzy logic based approach," in *Proceedings of the 11th International Conference Information Processing and Management of Uncertainty in Knowledge-based Systems*, 2006, pp. 2166–2172.
- [2] —, "On some types of linguistic summaries of time series," in *Proceedings of the 3rd International IEEE Conference "Intelligent Systems"*. IEEE Press, 2006, pp. 373–378.
- [3] —, "Linguistic summarization of time series using a fuzzy quantifier driven aggregation," *Fuzzy Sets and Systems*, vol. 159, no. 12, pp. 1485–1499, 2008.
- [4] J. Kacprzyk and A. Wilbik, "An extended, specificity based approach to linguistic summarization of time series," in *Proceedings of the 12th International Conference Information Processing and Management of Uncertainty in Knowledge-based Systems*, 2008, pp. 551–559.
- [5] —, "A new insight into the linguistic summarization of time series via a degree of support: Elimination of infrequent patterns," in *Soft Methods for Handling Variability and Imprecision*, D. Dubois, M. Lubiano, H. Prade, M. A. Gil, P. Grzegorzewski, and O. Hryniewicz, Eds. Springer-Verlag, Berlin and Heidelberg, 2008, pp. 393–400.
- [6] —, "Linguistic summarization of time series using linguistic quantifiers: augmenting the analysis by a degree of fuzziness," in *Proceedings of 2008 IEEE World Congress on Computational Intelligence*. IEEE Press, 2008, pp. 1146–1153.
- [7] Past performance does not predict future performance."
- [8] —, "www.personalfn.com/detail.asp?date=9/1/2007&story=3, Past performance is not everything."
- [9] —, "stockcasting.blogspot.com/2005/12/new-years-evepast-performance-is-no.html, new year's eve: past performance is no indication of future return."
- [10] R. Myers, "Using past performance to pick mutual funds," *Nation's Business*, 1997.
- [11] J. C. Bogle, *Common Sense on Mutual Funds: New Imperatives for the Intelligent Investor*. New York: Wiley, 1999.
- [12] investing: Look at more than a fund's past performance, U.S. Securities and Exchange Commission."
- [13] —, "www.personalfn.com/detail.asp?date=5/18/2007&story=2, A 10-step guide to evaluating mutual funds."
- [14] D. Ariely and D. Zakay, "A timely account of the role of duration in decision making," *Acta Psychologica*, vol. 108, no. 2, pp. 187–207, 2001.
- [15] H. Rachlin, *Judgment, Decision, and Choice: A Cognitive/Behavioral Synthesis*. W.H. Freeman & Company, 1989.
- [16] N. Liberman and Y. Trope, "The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory," *Journal of Personality and Social Psychology*, vol. 75, pp. 5–18, 1998.
- [17] J. Kacprzyk and A. Wilbik, "Temporal linguistic summaries of time series using fuzzy logic," in *Proceedings of IPMU2010 (in press)*, 2010.
- [18] J. Sklansky and V. Gonzalez, "Fast polygonal approximation of digitized curves," *Pattern Recognition*, vol. 12, no. 5, pp. 327–331, 1980.
- [19] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "An online algorithm for segmenting time series," in *Proceedings of the 2001 IEEE International Conference on Data Mining*, 2001.
- [20] —, "Segmenting time series: A survey and novel approach," in *Data Mining in Time Series Databases*, M. Last, A. Kandel, and H. Bunke, Eds. World Scientific Publishing, 2004.

- [21] R. R. Yager, "A new approach to the summarization of data," *Information Sciences*, vol. 28, pp. 69–86, 1982.
- [22] —, "On linguistic summaries in data," in *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. J. Frawley, Eds. MIT Press, Cambridge, USA, 1991, pp. 347–363.
- [23] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," *International Journal of General Systems*, vol. 30, pp. 33–154, 2001.
- [24] J. Kacprzyk, R. R. Yager, and S. Zadrożny, "A fuzzy logic based approach to linguistic summaries of databases," *International Journal of Applied Mathematics and Computer Science*, vol. 10, pp. 813–834, 2000.
- [25] —, "Fuzzy linguistic summaries of databases for an efficient business data analysis and decision support," in *Knowledge Discovery for Business Information Systems*, J. Z. W. Abramowicz, Ed. Boston: Kluwer, 2001, pp. 129–152.
- [26] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 9, no. 2, pp. 111–127, 1983.
- [27] J. Kacprzyk and S. Zadrożny, "Linguistic database summaries and their protoforms: toward natural language based knowledge discovery tools," *Information Sciences*, vol. 173, pp. 281–304, 2005.
- [28] J. Kacprzyk and A. Wilbik, "Linguistic summarization of time series using fuzzy logic with linguistic quantifiers: a truth and specificity based approach," in *Artificial Intelligence and Soft Computing ICAISC 2008*, L. Rutkowski, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds. Springer-Verlag, Berlin and Heidelberg, 2008, pp. 241–252.
- [29] —, "Towards an efficient generation of linguistic summaries of time series using a degree of focus," in *Proceedings of the 28th North American Fuzzy Information Processing Society Annual Conference – NAFIPS 2009*, 2009.
- [30] R. R. Yager, K. M. Ford, and A. J. Cañas, "An approach to the linguistic summarization of data," in *Uncertainty in Knowledge Bases, 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU '90, Paris, France, July 2-6, 1990, Proceedings*, B. Bouchon-Meunier, R. R. Yager, and L. A. Zadeh, Eds. Springer, 1990, pp. 456–468.
- [31] S. G. Sripada, E. Reiter, and I. Davy, "SumTime-Mousam: Configurable marine weather forecast generator," *Expert Update*, vol. 6, no. 3, pp. 4–10, 2003.
- [32] J. Kacprzyk and S. Zadrożny, "Data mining via protoform based linguistic summaries: Some possible relations to natural language generation," in *2009 IEEE Symposium Series on Computational Intelligence Proceedings*, Nashville, TN, 2009, pp. 217–224.
- [33] —, "Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries and natural language generation," *IEEE Transactions on Fuzzy Systems*, 2010, (forthcoming).
- [34] I. Batyrshin, "On granular derivatives and the solution of a granular initial value problem," *International Journal Applied Mathematics and Computer Science*, vol. 12, no. 3, pp. 403–410, 2002.
- [35] I. Batyrshin and L. Sheremetov, "Perception based functions in qualitative forecasting," in *Perception-based Data Mining and Decision Making in Economics and Finance*, I. Batyrshin, J. Kacprzyk, L. Sheremetov, and L. A. Zadeh, Eds. Springer-Verlag, Berlin and Heidelberg, 2006.
- [36] L. A. Zadeh, "A prototype-centered approach to adding deduction capabilities to search engines – the concept of a protoform," in *Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS 2002)*, 2002, pp. 523–525.
- [37] J. Kacprzyk, A. Wilbik, and S. Zadrożny, "Linguistic summarization of time series under different granulation of describing features," in *Rough Sets and Intelligent Systems Paradigms - RSEISP 2007*, M. Kryszykiewicz, J. F. Peters, H. Rybinski, and A. Skowron, Eds. Springer-Verlag, Berlin and Heidelberg, 2007, pp. 230–240.
- [38] R. R. Yager, "On measures of specificity," in *Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications*, O. Kaynak, L. A. Zadeh, B. Türksen, and I. J. Rudas, Eds. Springer-Verlag: Berlin, 1998, pp. 94–113.
- [39] —, "Measuring tranquility and anxiety in decision making: An application of fuzzy sets," *International Journal of General Systems*, vol. 8, pp. 139–146, 1982.
- [40] L. A. Zadeh, "Computation with imprecise probabilities," in *Proceedings of the 12th International Conference Information Processing and Management of Uncertainty in Knowledge-based Systems*, 2008.
- [41] J. Kacprzyk and A. Wilbik, "A multi-criteria evaluation of linguistic summaries of time series via a measure of informativeness," in *Proceedings of ICAISC2010 (in press)*, 2010.