LINGUISTIC SUMMARIES OF TIME SERIES:
ON SOME EXTENDED AGGREGATION TECHNIQUES

JANUSZ KACPRZYK
ANNA WILBIK
Systems Research Institute Polish Academy of Sciences

Summary

We further extend our approach to the linguistic summarization of time series (cf. Kacprzyk, Wilbik and Zadrozny) in which an approach based on a calculus of linguistically quantified propositions is employed, and the essence of the problem is equated with a linguistic quantifier driven aggregation of partial scores (trends). We proceed towards a multicriteria analysis of 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, 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.

Keywords: time series analysis, fuzzy logic, natural language, computing with words, linguistic summaries, time series summarization.

1. Introduction

Financial data analysis is one of the most important application areas of advanced data mining and knowledge discovery tools and techniques. For instance, in a report presented by Piatetsky-Shapiro (cf. http://www.kdnuggets.com) on top data mining applications in 2008, the first two positions are, in the sense of yearly increase:

- Investment/Stocks, up from 3% of respondents in 2007 to 14% of respondents in 2008 (350% increase),
- Finance, up from 7.2% in 2007 to 16.8% in 2008 (108% increase),

and this trend will presumably continue.

This paper is a follow up of our previous works (cf. Kacprzyk, Wilbik, Zadrozny [12–14] or Kacprzyk, Wilbik [8–11]) which deal with how to effectively and efficiently support a human decision maker in making decisions concerning investments in mutual funds.

Though decision makers are concerned with possible future gains/losses, and their decisions is related to the future, our aim is not the forecasting of the future daily prices. Instead, we follow a decision support paradigm, that is we try to provide the decision maker with some information that can be useful, not to replace the human decision maker.

For solving the problem, there may be two general approaches: first, to provide means to derive a price forecast for an investment unit so that the decision maker could “automatically” purchase what has been forecast. Unfortunately, the success has been much less than expected.

---

1 This work is partially supported by the Ministry of Science and Higher Education Grant no NN516 4309 33.
Basically, statistical methods just somehow extrapolate the past and do not use domain knowledge, intuition, inside information, etc. A natural solution may be to try to support the human decision maker by providing him/her with some additional useful information, while not getting involved in the very process of decision making.

From our perspective, the following philosophy will be followed. In all investment decisions the future is what really counts, and the past is irrelevant. But, the past is what we know, and the future is (completely) unknown. Behavior of the human being is to a large extend driven by his/her (already known) past experience. We usually assume that what happened in the past will also happen (to some, maybe large extent) in the future. This is basically, by the way, the very underlying assumption behind the statistical methods too!

This directly implies that the past can be employed to help the human decision maker. We present here a method to subsume the past, the past performance of an investment (mutual) fund, by presenting results in a very human consistent way, using natural language statements.

To start, in any information leaflet of an investment fund, there is a disclaimer stating that “Past performance is no indication of future returns” which is true. However, on the other hand, in a well known posting “Past Performance Does Not Predict Future Performance” [1], 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” [2], they state: “… disclaimers apart, as a practice investors continue to make investments based on a scheme’s 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”.

We can ask a natural question why it is so. Again, in a well known posting “New Year's Eve: Past performance is no indication of future return” [3], 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…”.

There are a multitude of similar statements in various well known postings, exemplified by Myers [25]: “…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… While some research has shown that consistently good performers continue to do well at a better rate than marginal performers, it also has shown a much stronger predictive value for consistently bad performers… Lousy performance in the past is indicative of lousy performance in the future…”.

And, further (cf. [29]): “While past performance does not necessarily predict future returns, it can tell you how volatile a fund has been”. And further, in the popular “A 10-step guide to evaluating mutual funds” [4], they say in the last advise: “Evaluate the fund’s performance. Every fund is benchmarked against an index like the BSE Sensex, Nifty, BSE 200 or the CNX 500 to cite a few names. Investors should compare fund performance over varying time frames vis-a-vis both the benchmark index and peers. Carefully evaluate the fund’s performance across market cycles particularly the downturns”. Therefore we think, that linguistic summaries of the past performers of an investment fund can be here a valuable tool as they may be easily understood by the humans as they are in natural language.
In 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/losses and their temporal evolution, e.g. some posting stress the role of absolute return mutual fund. McGowan says [24]: “Generally, mutual fund performance is compared relative to a benchmark. The relative return of a mutual fund measures how well a mutual fund has performed compared to its benchmark. Relative returns are important because it tells mutual fund investors whether or not they are getting what they paid for – returns in excess of the mutual fund’s benchmark… There is where absolute return mutual funds come into play. Absolute return mutual funds are managed with a specific return goal in mind. The goal of the absolute return mutual fund is to always have a positive return regardless of the market – and regardless of benchmarks.”

Here we extend our previous works on linguistic summarization of time series (cf. Kacprzyk, Wilbik, Zadrożyń [12–14] or Kacprzyk, Wilbik [8–11]), mainly towards a more complex evaluation of results. As the first step towards an intended comprehensive multicriteria 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 Yager’s [34] proposal on an informativeness measure of a linguistic summary which combines, via an appropriate aggregation operator, the degree of truth, focus and specificity.

One can also view this paper, as well as our other papers on this topic, from the viewpoint of natural language generation (NLG), a rapidly developing area (cf. Reiter and Dale [26] or Sripada et al. [28]), in its “numbers – to – words” direction, the essence of which is to devise tools and techniques to summarize a (large) set of numerical data by simple natural language statements comprehensible to the humans. A close relation between the linguistic summaries and NLG was pointed out in Kacprzyk and Zadrożyń [20, 21] who showed that the linguistic data summaries considered can be derived using an extended form of template based NLG systems, and also some simple phrase based NLG systems. This direction is very promising because one can use theoretical results of NLG, and also some available NLG software can be used.

2. Linguistic summaries of time series

In Yager’s basic approach [32, 33], used here, we have:
1. \( Y = \{y_1, y_2, \ldots, y_n\} \) is the set of objects (records) in the database \( D \), e.g., a set of employees; and
2. \( A = \{A_1, \ldots, A_m\} \) is the set of attributes (features) characterizing objects from \( Y \), e.g., a salary, age in the set of employees.

A linguistic summary of data comprise:
- 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” 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 \( R \), 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”).

\[
T(\text{most of employees are well-paid}) = 0.8
\]

\[
T(\text{most of young employees are well-paid}) = 0.85
\]

Thus, the core of a linguistic summary is a linguistically quantified proposition in the sense of Zadeh [35] which for (1) and (2) may be written as, respectively:

\[
Q_y \text{ s are } P
\]

\[
QR_y \text{ s are } P
\]

Then the truth (validity), \( T \), of a linguistic summary directly corresponds to the truth value of (3) and (4) which can be calculated using, for instance, the original Zadeh’s calculus of quantified propositions (cf. [35]).

First, in our approach we summarize the trends (segments) extracted from a time series, and we have to extract these segments assumed to be represented by a fragment of straight line. There are many algorithms for a piecewise linear segmentation of time series, including, e.g., on-line (sliding window) algorithms, bottom-up or top-down strategies (cf. Keogh [22, 23]). In our works [8–14] we used a simple on-line algorithm, a modification of Sklansky and Gonzalez [27].

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. Duration is the length of a single trend. 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). All of them are represented by a linguistic variables.

For practical reasons for all we use a fuzzy granularity (cf. Bathyrshin at al. [5–6]) to represent the values by a small set of linguistic labels as, e.g.: quickly increasing, increasing, slowly increasing, constant, slowly decreasing, decreasing, quickly decreasing. These values are equated with fuzzy sets.

For clarity and convenience, for dealing with linguistic summaries [19] we employ Zadeh’s [36] protoforms 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:

\[
\text{Among all segments, } Q \text{ are } P
\]

\[
\text{Among all } R \text{ segments, } Q \text{ are } P
\]

The protoforms are very convenient for various reasons, notably: they make it possible to devise general tools and techniques for dealing with a variety of statements concerning different domains and problems, and their form is often easily comprehensible to domain specialists.
The quality of linguistic summaries can be evaluated in many different ways, e.g., using the degree of truth, specificity, appropriateness or others (cf. Kacprzyk, Yager [15], Kacprzyk, Yager and Zadroży[16, 17] or Kacprzyk and Zadroży[18]). Yager [34] 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.

2.1. Truth value

The truth value (a degree of truth or validity), introduced by Yager in [32], 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 [], the truth value is calculated as:

\[
T(\text{Among all } y's, Q \text{ are } P) = \mu_0 \left( \frac{1}{n} \sum_{i=1}^{n} \mu_P(y_i) \right)
\]

where \(a \land b = \min(a, b)\) (more generally, a t-norm, cf. Kacprzyk, Wilbik and Zadroży[13]).

Zadeh’s calculus of linguistically quantified propositions is known to perform poorly in some cases, notably for small data sets, and some other approaches for handling linguistic quantifiers are known (cf. Glöckner [7]) which do not exhibit such a deficiency. However, Zadeh’s approach has proved to be implementable, is simple, and can more easily deal with protoforms. These virtues are relevant for our application, and hence Zadeh’s method is used.

2.2. 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 [30].

We will consider the original Yagers proposal [30], in which specificity measures the degree to which a fuzzy subset contains one and only one element. The measure of specificity is a measure \(Sp : \mathbb{I}^I \rightarrow I, \quad I \in [0,1]\) 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. \(\frac{\partial Sp(a)}{\partial a} > 0\) and \(\frac{\partial^2 Sp(a)}{\partial a^2} \leq 0\) for all \(j \leq 2\).

In [31] Yager proposed a measure of specificity as

\[
Sp(A) = \int_0^{\text{area}} \frac{1}{\text{card}(A_\alpha)} d\alpha
\]
where $\alpha_{\text{max}}$ is the largest membership grade in $A$, $A_\alpha$ is the $\alpha$-level set of $A$, (i.e. $A_\alpha = \{ x: A(x) \geq \alpha \}$) and $\text{card}(A_\alpha)$ is the number of elements in $A_\alpha$.

![Trapezoidal membership function of a set.](image)

**Fig. 1. 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 [37]. Moreover, they can be very easily interpreted and defined by a user not familiar with fuzzy sets and logic, as in Figure 1. 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 coarseness and computational complexity. In such a case measure of specificity of a fuzzy set $A$

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

(10)

### 2.3. Degree of focus

The very purpose of a degree of focus is to limit the search for best linguistic summaries by taking into account some additional information in addition to the degree of truth (validity). The extended form of linguistic summaries (6) 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 degree of focus measures how many trends fulfill property $R$. That is, we focus our attention on such trends, fulfilling property $R$. The degree of focus, only for (6) is calculated as:

$$d_{fo}\left(\text{Among all } R_\text{y}'s, Q \text{ are } P\right) = \mu_{\theta}\left(\frac{1}{n} \sum_{i=1}^{n} \mu_{R}(y_i)\right)$$

(11)

It provides a measure that, in addition to the degree of truth, can help control the process of discarding nonpromising linguistic summaries; cf. Kacprzyk and Wilbik [11]. 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.

### 2.4. Measure of informativeness

The idea of the measure of informativeness (cf. Yager [34]) 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.

Measure of informativeness depends on the type of quantifier. Note that we consider here regular and monotonic quantifiers. For monotonically non-decreasing quantifier (like “most”, “almost all”, etc.) Yager [34] proposed the following measure of informativeness of a simple form summary:

\[
I(\text{Among all } y\text{'s, } Q \text{ are } P) = (T \cdot \text{Sp}(Q) \cdot \text{Sp}(P)) \lor ((1 - T) \cdot \text{Sp}(\hat{Q}) \cdot \text{Sp}(P^c))
\]

(12)

where \( P^c \) is the negation of summarizer \( P \), i.e. \( \mu_{P_c}(\cdot) = 1 - \mu_P(\cdot) \) and \( \hat{Q} \) is the antonym of \( Q \), \( \hat{Q}(x) = Q(1 - x) \). \( Q^c \) is the negation of antonym of \( Q \), i.e. \( \mu_{Q_c}(x) = 1 - \mu_Q(1 - x) \). \( \text{Sp}(Q) \) is specificity of \( Q \) defined as in subsection 3.2, similarly it is calculated for \( \hat{Q}^c \), \( P \) and \( P^c \).

For the extended form summary we propose the following measure

\[
I(\text{Among all } y\text{'s, } Q \text{ are } P) = (T \cdot \text{Sp}(Q) \cdot \text{Sp}(P) \cdot \text{Sp}(R) \cdot d_{\text{foc}}) \lor
((1 - T) \cdot \text{Sp}(\hat{Q}) \cdot \text{Sp}(P^c) \cdot \text{Sp}(R) \cdot d_{\text{foc}})
\]

(13)

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

In case of non-increasing quantifiers such as “minority” or “less than 20%”, Yager [34] proposes the following measure of informativeness of a simple form summary:

\[
I(\text{Among all } y\text{'s, } Q \text{ are } P) = (T \cdot \text{Sp}(Q) \cdot \text{Sp}(P)) \lor ((1 - T) \cdot \text{Sp}(Q^c) \cdot \text{Sp}(P^c))
\]

(14)

where \( Q^c \) is the negation of quantifier \( Q \), i.e. \( \mu_{Q^c}(\cdot) = 1 - \mu_Q(\cdot) \).

For the extended form summary we propose the following measure

\[
I(\text{Among all } y\text{'s, } Q \text{ are } P) = (T \cdot \text{Sp}(Q) \cdot \text{Sp}(P) \cdot \text{Sp}(R) \cdot d_{\text{foc}}) \lor
((1 - T) \cdot \text{Sp}(Q^c) \cdot \text{Sp}(P^c) \cdot \text{Sp}(R) \cdot d_{\text{foc}})
\]

(15)

Here in those formulas different values are aggregated by the product. We could think to use instead of the product other t-norms. However, for example, the minimum would ignore other values, 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.
3. 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 2 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.

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 summaries are presented in Table 1. They are ordered according to the truth value.

The summaries in the table are ordered according to the truth value, and then by the degree of focus. Generally, the simple form summaries, (e.g. Among all y’s, most are short) have a higher measure of informativeness, as they describe whole data set. The measure of informativeness of the extended form summaries is smaller, because they describe only a subset of the data.

This measure considers also number and quality of the adjectives used. For instance, for “Among all decreasing y’s, almost all are short”, with $I = 0.1166$, and “Among all decreasing y’s, most are short and low”, with $I = 0.1333$, the latter, although it has a bit smaller truth value, is more informative, as it provides additional information. This is a more general property resulting from our experiments.

It seems that the measure of informativeness is a good evaluation of the amount of information carried by the summary. Moreover, as it combines the measure of truth, focus and specificity in a intuitively appealing yet simple way, may be viewed as an effective and efficient tools for a multi-criteria assessment of linguistic summaries of times series.

![Mutual fund quotations](image)

*Fig. 2. Daily quotations of an investment (mutual) fund in question*
Tab. 1. Results obtained for 3 labels only for each attribute

<table>
<thead>
<tr>
<th>Linguistic Summary</th>
<th>T</th>
<th>$d_foc$</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among all low y’s, most are short</td>
<td>1.0000</td>
<td>0.7560</td>
<td>0.2736</td>
</tr>
<tr>
<td>Among all decreasing y’s, almost all are short</td>
<td>1.0000</td>
<td>0.2720</td>
<td>0.1166</td>
</tr>
<tr>
<td>Among all increasing y’s, almost all are short</td>
<td>1.0000</td>
<td>0.2668</td>
<td>0.1143</td>
</tr>
<tr>
<td>Among all short and increasing y’s, most are low</td>
<td>1.0000</td>
<td>0.2483</td>
<td>0.1444</td>
</tr>
<tr>
<td>Among all decreasing y’s, most are low</td>
<td>0.9976</td>
<td>0.2720</td>
<td>0.0596</td>
</tr>
<tr>
<td>Among all short and decreasing y’s, most are low</td>
<td>0.9969</td>
<td>0.2645</td>
<td>0.1533</td>
</tr>
<tr>
<td>Among all increasing y’s, most are short and low</td>
<td>0.9860</td>
<td>0.2668</td>
<td>0.1352</td>
</tr>
<tr>
<td>Among all y’s, most are short</td>
<td>0.9694</td>
<td>–</td>
<td>0.5012</td>
</tr>
<tr>
<td>Among all decreasing y’s, most are short and low</td>
<td>0.9528</td>
<td>0.2720</td>
<td>0.1333</td>
</tr>
<tr>
<td>Among all y’s, most are low</td>
<td>0.9121</td>
<td>–</td>
<td>0.3512</td>
</tr>
<tr>
<td>Among all short and constant y’s, most are low</td>
<td>0.8408</td>
<td>0.2741</td>
<td>0.1597</td>
</tr>
<tr>
<td>Among all moderate y’s, most are short</td>
<td>0.8274</td>
<td>0.2413</td>
<td>0.0619</td>
</tr>
<tr>
<td>Among all constant y’s, most are low</td>
<td>0.8116</td>
<td>0.4612</td>
<td>0.1239</td>
</tr>
<tr>
<td>Among all medium and constant y’s, most are low</td>
<td>0.7646</td>
<td>0.1265</td>
<td>0.0650</td>
</tr>
<tr>
<td>Among all medium y’s, most are low</td>
<td>0.7167</td>
<td>0.1524</td>
<td>0.0372</td>
</tr>
</tbody>
</table>

4. Conclusions

We extended our approach to the linguistic summarization of time series towards a multi-criteria analysis of 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.

5. Literature


LINGWISTYCZNE PODSUMOWANIA CIĄGÓW CZASOWYCH:
PEWNE ROZSZERZENIA TECHNIK AGREGACJI

Streszczenie

W artykule przedstawiono kolejne rozszerzenie poprzednich prac autorów (Kacprzyk, Wilbik, Zadrozny), dotyczących podsumowań lingwistycznych, w których podstawowym podejściem jest rachunek lingwistycznie kwantyfikowanych wyrażeń, a głównym zagadnieniem jest oparte na kwantyfikatorach lingwistycznych agregowanie częściowych ocen (trendów). Wykorzystuje się podejście wielokryterialne do podsumowań z użyciem miary informatywności, zaproponowanej przez Yagera, odwołującą się do kryteriów prawdy, zogniskowania i specyficzności. Tego rodzaju podejście wielokryterialne wydaje się zarówno efektywne merytorycznie, jak i wystarczająco zrozumiałe i intuicyjne. Przytoczono obszerne przykłady dla notowań funduszy inwestycyjnych.

Słowa kluczowe: analiza ciągów czasowych, logika rozmyta, język naturalny, obliczenia werbalne, podsumowania lingwistyczne, podsumowania ciągów czasowych.

Janusz Kacprzyk
Anna Wilbik
Instytut Badań Systemowych PAN
Newelska 6, 01-447 Warszawa
e-mail: Janusz.Kacprzyk@ibspan.waw.pl
Anna.Wilbik@ibspan.waw.pl