2009 Ninth International Conference on Intelligent Systems Design and Applications

# Linguistic summaries of time series using a degree of appropriateness as a measure of interestingness

Janusz Kacprzyk[12], Anna Wilbik[1]
[1] Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
kacprzyk@ibspan.waw.pl
wilbik@ibspan.waw.pl
[2] WIT – Warsaw School of Information Technology
Newelska 6, 01-447 Warsaw, Poland

## Abstract

*We further extend our approach to the linguistic summarization of time series (cf. Kacprzyk, Wilbik and Zadrożny [9, 10, 11, 12]) 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). In addition to the basic criterion of a degree of truth (validity), we also use as a degree of appropriateness as an additional quality criterion. However, for simplicity and tractability, we use in the first shot the degrees of truth (validity) and focus, which usually reduce the space of possible linguistic summaries to a considerable extent, and then – for a usually much smaller set of linguistic summaries obtained – we use the degree of appropriateness to make a final choice as it gives us an additional quality of being able to detect how surprising, i.e. valuable, a linguistic summary obtained is. We also mention relations to natural language generation (NLG) as pointed out recently by Kacprzyk and Zadrożny [19]. We show an application to the absolute performance type analysis of daily quotations of an investment fund, and the numerical results are promising. The linguistic summaries obtained using this additional quality criterion of a degree of appropriateness seem to better reflect human intents and interest.*

## 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 form 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, Zadrożny [9, 10, 11, 12] or Kacprzyk, Wilbik [7]) 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. 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

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.227

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 vary 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" [2], 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" [3], 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".

There are a multitude of similar statements in various well known postings, exemplified by Myers [22]: "...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. [24]): "While past performance does not necessarily predict future returns, it can tell you how volatile a fund has been". And furthr, in the popular "A 10-step guide to evaluating mutual funds" [1], they say in the last advise: "Evaluate the funds 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 funds 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.

Here we extend our previous works on linguistic summarization of time series (cf. Kacprzyk, Wilbik, Zadrożny [9, 11, 12] or Kacprzyk, Wilbik [7]), mainly towards a more complex evaluation of results. The basic criterion for evaluation linguistic summaries is a degree of truth (cf. our papers [9, 11, 13]). However, later Kacprzyk and Yager [14] and Kacprzyk, Yager and Zadrożny [15, 16] and Kacprzyk and Zadrożny [18, 17] introduced additional quality criteria, notably a degree of appropriateness, which will be discussed here.

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 [23]), 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 Zadrożny [19] 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. It will be explored in a later paper.

## 2. Linguistic summaries of time series

In Yager's basic approach [26], 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) $A = \{A_1, A_2, \ldots, A_m\}$ is the set of attributes (features) characterizing objects from $Y$, e.g., a salary, age in the set of employees.

A linguistic summary includes: (1) 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*); (2) a quantity in agreement $Q$, i.e. a linguistic quantifier (e.g. *most*); (3) truth (validity) $\mathcal{T}$ of the summary, i.e. a number from the interval $[0, 1]$ assessing the truth (validity) of the summary (e.g. 0.7); and optionally, (4) a qualifier $R$, i.e. another attribute together with a linguistic value (fuzzy predicate) defined on the domain of attribute $A_k$ determining a (fuzzy) subset of $Y$ (e.g. *young* for attribute *age*).

Thus, a linguistic summary in a simple or extended form (including a qualifier, e.g. *young*, may be exemplified by, respectively:

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

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

Thus, the core of a linguistic summary is a linguistically quantified proposition in the sense of Zadeh [27] which for

(1) and and (2) may be written as, respectively:

$$Qy's \text{ are } P \tag{3}$$

$$QRy's \text{ are } P \tag{4}$$

Then the truth (validity), $\mathcal{T}$, 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. [27]).

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 [20, 21]). In our works [9, 13, 7] we used a simple on-line algorithm, a modification of Sklansky and Gonzalez [25] which give good results.

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 granulation (cf. Bathyrshin at al. [4, 5]) 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 [18] we employ Zadeh's [28] 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:

$$\text{Among all segments, } Q \text{ are } P \tag{5}$$

$$\text{Among all } R \text{ segments, } Q \text{ are } P \tag{6}$$

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 linguistic summary is evaluated based on the truth value. We compute it while we generate summaries, and hence we may eliminate some summaries with small truth value. We introduce a degree of focus to further limit the search space of all possible extended form summaries.

## 2.1. Truth value

The truth value (a degree of truth or validity), introduced by Yager in [26], 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 [27], the truth value is calculated as:

$$\mathcal{T}(\text{Among } y\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_P(y_i)\right) \tag{7}$$

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

where $a \wedge b = \min(a, b)$ (more generally, a $t$-norm).

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 [6] 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 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:

$$d_{foc}(\text{Among all } Ry\text{, 's } Q \text{ are } P) = \frac{1}{n}\sum_{i=1}^{n}\mu_R(y_i) \tag{9}$$

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 [8]. We tacitly assume that the degree of focus is an obvious additional criterion, and will not mention it as such, so that the degree of appropriateness will only be mentioned as an additional criterion.

### 2.3. Degree of appropriateness

The degree of appropriateness, introduced by Kacprzyk and Yager [14], and Kacprzyk, Yager and Zadrożny [15, 16] indicates if, and to which degree, the obtained summary is surprising to us. It is believed to be one of the most relevant measure of the summary. In our case of linguistic summaries of time series (trends) it is calculated as

$$d_a = \left| \prod_{i=1}^{m} \frac{card\{y : \mu_{A_i}(y) > 0\}}{n} - \frac{card\{y : \forall A_i \ \mu_{A_i}(y) > 0\}}{n} \right| \quad (10)$$

where $A_i$ is a predicate corresponding to summarizer $P$ or additionally in the case of extended summaries qualifier $R$. Note, that this measure equals 0 in simple form summary with only one predicate.

We can interpret this value as follows. Let us assume, that we have $n$ trends, and we consider 2 properties $A$ and $B$. Let us assume, that 50% of trends have property $A$ and 50% have property $B$. Assuming that those properties are independent, we expect that 25% of trends have both of the properties. This value is calculated as the first expression in the difference. The exact number of object having those two properties is calculated as the second expression in the above formula. If the real proportion of trends having those both properties is much higher or lower than expected, we will find this summary interesting.

The maximum value of this measure depends on the number of properties in the summary, and it can be calculated as $\sqrt[m-1]{\frac{1}{m}}$, where $m$ is the number of properties in the summary. In order to compare those values for summaries with a different number of predicates we should normalize by dividing the obtained value of degree of appropriateness $d_a$ by the maximum value of this measure for a given number of predicates in the summary.

This measure is similar to the well known Piatetsky-Shapiro interest function for association rules. It is used to quantify the correlation between attributes in a simple classification rule.

## 3. Numerical results

The method proposed in this paper was tested on data on quotations of an investment (mutual) fund that invests at most 50% of assets in shares listed at the Warsaw Stock Exchange. Data shown in Figure 1 were collected from January 2002 until March 2009 with the value of one share equal to PLN 12.06 in the beginning of the period to PLN 21.82 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.

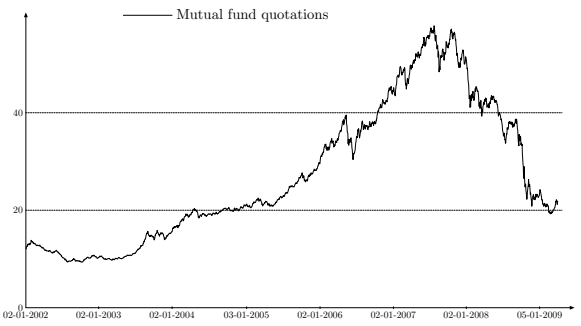


**Figure 1. Mutual fund quotations**

It should be noted that the example shown below is meant to illustrate the method proposed by analyzing the absolute performance of a given investment fund. We do not deal here with a more common way of analyzing an investment fund by relating its performance to a benchmark (or benchmarks) exemplified by an average performance of a group of (similar) funds, a stock market index or a synthetic index reflecting, for instance, the bond versus stock allocation.

Using the modified Sklansky and Gonzalez algorithm (cf. [25]) and $\varepsilon = 0.25$, which was experimentally found to correspond to the best granularity level, and we obtained 422 extracted trends. The shortest trend took 1 time unit (day) only, while the longest one – 71. Clearly, the handling of trend with such a varying length is not trivial conceptually and numerically but we will not consider this issue here due to space limitation.

We have applied different granulations, namely with 3 and 5 labels for each feature (dynamics of change, duration and variability). Minimal accepted truth value was 0.7 and the degree of focus threshold was 0.1. The degree of focus, and the method of effective and efficient generating summaries is described in Kacprzyk and Wilbik's paper [8].

If we have used 3 labels for dynamics of change (decreasing, constant and increasing), 3 labels for duration (short, medium length and long) and 3 labels for variability (low, moderate and high) we have obtained the summaries shown in Table 1.

Note that the following 3 summaries:

- Among all decreasing $y$'s, most are short, $d_a$=0.0112
- Among all decreasing $y$'s, most are short and low, $d_a$=0.0234
- Among all decreasing $y$'s, most are low, $d_a$=0.0057

have the same truth values and the degrees of focus are equal. However their degrees of appropriateness are different. From this perspective the second summary mentioned here seems to be most interesting.

If we used 5 labels for dynamics of change (quickly decreasing, decreasing, constant, increasing, and quickly increasing), 5 labels for duration (very short, short, medium

**Table 1. Results for 3 labels**

| linguistic summary | $\mathcal{T}$ | $d_{foc}$ | $d_a$ |
|---|---|---|---|
| Among all low $y$'s, most are short | 1 | 0.7227 | 0.0043 |
| Among all increasing $y$'s, most are short | 1 | 0.2984 | 0.0105 |
| Among all decreasing $y$'s, most are short | 1 | 0.2880 | 0.0112 |
| Among all decreasing $y$'s, most are short and low | 1 | 0.2880 | 0.0234 |
| Among all decreasing $y$'s, most are low | 1 | 0.2880 | 0.0057 |
| Among all short and decreasing $y$'s, most are low | 1 | 0.2842 | 0.0234 |
| Among all medium $y$'s, most are constant | 1 | 0.1308 | 0.0194 |
| Among all $y$'s, most are short | 1 | | 0 |
| Among all increasing $y$'s, most are low | 0.9610 | 0.2984 | 0.0016 |
| Among all short and increasing $y$'s, most are low | 0.9588 | 0.2946 | 0.0160 |
| Among all short $y$'s, most are low | 0.9483 | 0.8341 | 0.0043 |
| Among all increasing $y$'s, most are short and low | 0.9386 | 0.2984 | 0.0160 |
| Among all $y$'s, most are low | 0.8455 | | 0 |
| Among all moderate $y$'s, most are short | 0.7393 | 0.2483 | 0.0143 |
| Among all short and constant $y$'s, most are low | 0.7325 | 0.2565 | 0.0330 |
| Among all moderate $y$'s, most are constant | 0.7024 | 0.2483 | 0.0206 |

**Table 2. Results for 5 labels**

| linguistic summary | $\mathcal{T}$ | $d_{foc}$ | $d_a$ |
|---|---|---|---|
| Among all v. short $y$'s, most are v. low | 1 | 0.7180 | 0.0252 |
| Among all v. low $y$'s, most are v. short | 1 | 0.6141 | 0.0252 |
| Among all increasing $y$'s, most are v. short | 1 | 0.1903 | 0.0079 |
| Among all q. decreasing $y$'s, most are v. short | 1 | 0.1484 | 0.0096 |
| Among all q. decreasing $y$'s, most are v. short and v. low | 1 | 0.1484 | 0.0259 |
| Among all q. decreasing $y$'s, most are v. low | 1 | 0.1484 | 0.0102 |
| Among all v. short and q. decreasing $y$'s, most are v. low | 1 | 0.1464 | 0.0259 |
| Among all decreasing $y$'s, most are v. short | 1 | 0.1434 | 0.0088 |
| Among all v. short and decreasing $y$'s, most are v. low | 1 | 0.1275 | 0.0173 |
| Among all q. increasing $y$'s, most are v. short | 1 | 0.1101 | 0.0063 |
| Among all q. increasing $y$'s, most are v. short and v. low | 1 | 0.1101 | 0.0199 |
| Among all q. increasing $y$'s, most are v. low | 1 | 0.1101 | 0.0080 |
| Among all v. short and q. increasing $y$'s, most are v. low | 1 | 0.1100 | 0.0199 |
| Among all short $y$'s, most are constant | 0.8999 | 0.1979 | 0.0174 |
| Among all low $y$'s, most are constant | 0.8872 | 0.1471 | 0.0157 |
| Among all decreasing $y$'s, most are v. low | 0.8585 | 0.1434 | 0.0042 |
| Among all decreasing $y$'s, most are v. short and v. low | 0.8477 | 0.1434 | 0.0174 |
| Among all $y$'s, most are v. short | 0.8360 | | 0 |
| Among all v. short and increasing $y$'s, most are v. low | 0.7720 | 0.1611 | 0.0115 |
| Among all v. short and constant $y$'s, most are v. low | 0.7573 | 0.1857 | 0.0240 |

length, long, and very long) and 5 labels for variability (very low, low, moderate, high, and very high) we have obtained the summaries shown in Table 2.

These results have been found useful by our collaborating domain experts.

## 4. Conclusions

We further extended our approach to the linguistic summarization of time series 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). In addition to the standard quality criterion, which is the degree of truth, augmented with the degree of focus for truncating non-promising linguistic summaries, we use as an additional criterion the degree of appropriateness. It gives us an additional quality of being able to detect how surprising, i.e. valuable, a linguistic summary obtained is. Moreover, we have mentioned that a fruitful area of research may be to use a recent results of Kacprzyk and Zadrożny [19] who have indicated close relations between the employed approach to linguistic data summaries and natural language generation (NLG), notably to some extended template base and some simple phrase based NLG systems.

There are many relevant issues related to the approach presented like the sensistivity of the results obtained to the assumed levels of granulation (in the segmentation of time series, number of possible linguiostic descriptions of various paparemeters, etc.), the choice of quantifiers, the generation of best summaries, etc. but these issues cannot be dealt with due to space limitation. We showed an application to the absolute performance type analysis of daily quotations

of an investment fund, and the numerical results are promising. The linguistic summaries obtained using this additional quality criterion of a degree of appropriateness seem to better reflect human intents and interest.

## References


[1] A 10-step guide to evaluating mutual funds. www.personalfn.com/detail.asp?date=5/18/2007&story=2.

[2] Past performance does not predict future performance. www.freemoneyfinance.com/2007/01/past_performanc.html.

[3] Past performance is not everything. www.personalfn.com/detail.asp? date=9/1/2007&story=3.

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

[5] I. Batyrshin and L. Sheremetov. Perception based functions in qualitative forecasting. In I. Batyrshin, J. Kacprzyk, L. Sheremetov, and L. A. Zadeh, editors, *Perception-based Data Mining and Decision Making in Economics and Finance*. Springer-Verlag, Berlin and Heidelberg, 2006.

[6] I. Glöckner. *Fuzzy Quantifiers, A Computational Theory*, volume 193. Springer-Verlag, Berlin and Heidelberg, 2006.

[7] 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*, pages 551–559, 2008.

[8] J. Kacprzyk and A. Wilbik. 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.

[9] J. Kacprzyk, A. Wilbik, and S. Zadrożny. 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*, pages 2166–2172, 2006.

[10] J. Kacprzyk, A. Wilbik, and S. Zadrożny. Analysis of time series via their linguistic summarization: the use of the sugeno integral. In *Proceedings of the 7th International Conference on Intelligent Systems Design and Applications - ISDA 2007*, pages 262–267. IEEE Press, 2007.

[11] J. Kacprzyk, A. Wilbik, and S. Zadrożny. Linguistic summaries of time series via an owa operator based aggregation of partial trends. In *Proceedings of the FUZZ-IEEE 2007 IEEE International Conference on Fuzzy Systems*, pages 467–472. IEEE Press, 2007.

[12] J. Kacprzyk, A. Wilbik, and S. Zadrożny. On linguistic summaries of time series via a quantifier based aggregation using the sugeno integral. In O. Castillo, P. Melin, J. Kacprzyk, and W. Pedrycz, editors, *Hybrid Intelligent Systems Analysis and Design*, pages 421–439. Springer-Verlag, Berlin and Heidelberg, 2007.

[13] J. Kacprzyk, A. Wilbik, and S. Zadrożny. Linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Fuzzy Sets and Systems*, 159(12):1485–1499, 2008.

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

[15] 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*, 10:813–834, 2000.

[16] J. Kacprzyk, R. R. Yager, and S. Zadrożny. Fuzzy linguistic summaries of databases for an efficient business data analysis and decision support. In J. Z. W. Abramowicz, editor, *Knowledge Discovery for Business Information Systems*, pages 129–152. Kluwer, Boston, 2001.

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

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

[19] 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*, pages 217–224, Nashville, TN, 2009.

[20] 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.

[21] E. Keogh, S. Chu, D. Hart, and M. Pazzani. Segmenting time series: A survey and novel approach. In M. Last, A. Kandel, and H. Bunke, editors, *Data Mining in Time Series Databases*. World Scientific Publishing, 2004.

[22] R. Myers. Using past performance to pick mutual funds. *Nation's Business*, Oct, 1997. findarticles.com/p/articles/mi_m1154/is_n10_v85/ai_19856416.

[23] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2006.

[24] U. Securities and E. Commission. Mutual fund investing: Look at more than a fund's past performance. www.sec.gov/investor/pubs/mfperform.htm.

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

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

[27] L. A. Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems*, 9(2):111–127, 1983.

[28] 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)*, pages 523–525, 2002.