122/2009

Raport Badawczy

Research Report

RB/53/2009

A multi-criteria evaluation of linguistic summaries of time series via a measure of informativeness

J. Kacprzyk, A. Wilbik

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Pracowni zgłaszający pracę:
Prof. dr hab. inż. Janusz Kacprzyk

Warszawa 2009

A multi-criteria evaluation of linguistic summaries of time series via a measure of informativeness

Janusz Kacprzyk **Anna Wilbik**

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 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.

1 Introduction

This paper is an extension of the our previous works (cf. Kacprzyk, Wilbik and Zadrożny [12], and Kacprzyk and Wilbik [7, 8, 9, 10], Kacprzyk and Zadrożny [17, 18]) in which fuzzy logic, computing with words, and natural language generation were employed to derive a linguistic summary of a time series in the sense of a verbalization of how the time series behaves in regards to both the temporal evolution of values, their variability, etc.

This is a different approach to the analysis of time series than the usual forecasting/prediction analyses, and is rather focused on providing tools and techniques for supporting decision making by a human analyst. The use of linguistic summaries is an example of verbalization of the results of data analysis which is less common and popular than visualization.

First, one should notice that in virtually all non-trivial practical problems, notably in finance (to be more specific, investments in mutual funds considered by us) a decision support paradigm is employed, i.e. a decision is made by a human analyst based on some results of data analysis, modeling, calculations, etc. provided by the system. Here we consider some verbal summary of the past, with respect to the time series (more specifically, quotations of a mutual fund), as additional information that may be of much use to the analyst.

There is an ample rationale for this approach. On the one hand, in any mutual fund information leaflet, there is a disclaimer like "Past performance is no indication of future returns", which is true. However, on the other hand (cf. "Past Performance Does Not Predict Future Performance" [2]), they also state: "... 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". Similarly, in "Past performance is not everything" [3], there is "... 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". Moreover, in "New Year's Eve: Past performance is no indication of future return" [1], 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 ...", or in [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: in [6], we have: "... 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 [23], we find: "While past performance does not necessarily predict future returns, it can tell you how volatile a fund has been". There are a multitude of similar opinions expressed by top investment theorists, practitioners and advisors.

We use a slightly unorthodox approach to the summarization of the past performance of an investment fund by using a natural language summary exemplified by "most of long trends are slightly increasing". We use Yagers [27, 28] approach to linguistic summarization of numerical data that is based on fuzzy logic, more specifically on a calculus of linguistically quantified propositions. An important new directions, initiated by Kacprzyk and Zadrożny [19] is here a suggestion that a proper setting in which to derive linguistic data summaries may be within natural language generation (NLG), a modern, rapidly developing field of computer science and computational linguistics. This will not be discussed in more details here.

The analysis of time series data involves different elements but we concentrate on the specifics of our approach. First, we need 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, 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. [24]) though many other methods can be used cf. Keogh et al. [20, 21].

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 [27, 28] and then shown more profoundly and in an implementable way in Kacprzyk and Yager [13], and Kacprzyk, Yager and Zadrożny [14, 15], that the most comprehensive and meaningful will be a linguistic quantifier driven aggregation resulting in linguistic summaries exemplified by "Most trends are short" or "Most long trends are increasing" which 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 [31] protoforms as proposed by Kacprzyk and Zadrożny [18].

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 indicator. Some other indicators like a degree of specificity, focus, fuzziness, etc. have also been proposed by Kacprzyk and Wilbik [7, 8, 9, 10]. 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 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 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 Data[1] in which English summary descriptions of a time series data set are sought by using advanced time series and NLG (natural language generation) technologies [25]. 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 Zadrożny [19, 16].

2 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 [27]. 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) T of the summary and optionally, (4) a qualifier R (e.g.

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

young for attribute *age*).

Thus, a linguistic summary may be exemplified by

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

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

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

Thus, basically the core of a linguistic summary is a linguistically quantified proposition in the sense of Zadeh [31] which may be written, respectively as

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

3 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 [20, 21]).

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. [4, 5]) 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 [32] protoforms for dealing with linguistic summaries [18]. 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:

$$\text{Among all segments, } Q \text{ are } P \tag{4}$$

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

– an extended form:

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

e.g.: "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.

3.1 Truth value

The truth value (a degree of truth or validity), introduced by Yager in [27], 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 [31] 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:

$$\mathcal{T}(\text{Among all } y\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_P(y_i)\right) \tag{6}$$

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

where $\wedge$ is the minimum operation (more generally it can be another appropriate operator, notably a t-norm). In Kacprzyk, Wilbik and Zadrozny [11] 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.

3.2 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 [29].

We will consider the original Yagers proposal [29], in which specificity measures the degree to which a fuzzy subset contains one and only one element. The measure of specificity is a measure $Sp : I^X \longrightarrow I$, $I \in [0, 1]$ if it has the following properties: (1)

$Sp(A) = 1$ if and only if $A = \{x\}$, (is a singleton set), (2) $Sp(\varnothing) = 0$, and (3) $\frac{\partial Sp(A)}{\partial a_1} > 0$ and $\frac{\partial Sp(A)}{\partial a_j} \leq 0$ for all $j \geq 2$.

In [26] Yager proposed a measure of specificity as

$$Sp(A) = \int_0^{\alpha_{max}} \frac{1}{card(A_\alpha)} d\alpha \tag{8}$$

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

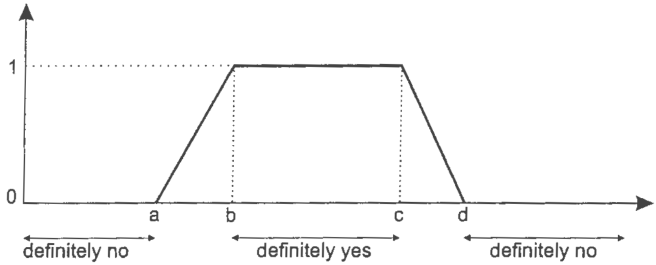

Figure 1: A trapezoidal membership function of a set

In our summaries to define the membership functions of the linguistic values we use trapezoidal functions, as they are sufficient in most applications [33]. 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 cointension and computational complexity. In such a case measure of specificity of a fuzzy set A

$$Sp(A) = 1 - \frac{c + d - (a + b)}{2} \tag{9}$$

3.3 Degree of focus

The extended form of linguistic summaries (5) 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 [10] is to give the proportion of trends satisfying property R to all trends extracted from the time series.

The degree of focus is similar in spirit to a degree of covering [9], however it measures how many trends fulfill property R. That is, we focus our attention on such trends, fulfilling property R. The degree of focus makes obviously sense for the extended form summaries only, and is calculated as:

$$d_f(\text{Among all } Ry\text{'s } Q \text{ are } P) = \frac{1}{n}\sum_{i=1}^{n}\mu_R(y_i) \tag{10}$$

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.

3.4 Measure of informativeness

The idea of the measure of informativeness(cf. Yager [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 [30] proposed the following measure of informativeness of a simple form summary

$$
\begin{aligned}
&I(\text{Among all } y\text{'s } Q \text{ are } P) = \\
&= \; (\mathcal{T} \cdot Sp(Q) \cdot Sp(P)) \vee ((1-\mathcal{T}) \cdot Sp(Q^c) \cdot Sp(P^c)) \qquad (11)
\end{aligned}
$$

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 defined as in subsection 3.2, similarly it is calculated for Q^c, P and P^c.

For the extended form summary we propose the following measure

$$
\begin{aligned}
I(\text{Among all } Ry\text{'s } Q \text{ are } P) \; &= \; (\mathcal{T} \cdot Sp(Q) \cdot Sp(P) \cdot Sp(R) \cdot d_f) \\
&\vee ((1-\mathcal{T}) \cdot Sp(Q^c) \cdot Sp(P^c) \cdot Sp(R) \cdot d_f) \qquad (12)
\end{aligned}
$$

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

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 resuklts from, for instance, decision analysis.

4 Numerical results

The method proposed in this paper 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 Fig. 2 were collected from January 2002 until the end of October 2009 with the value of one share equal to PLN 12.06 in the beginning of the period to PLN

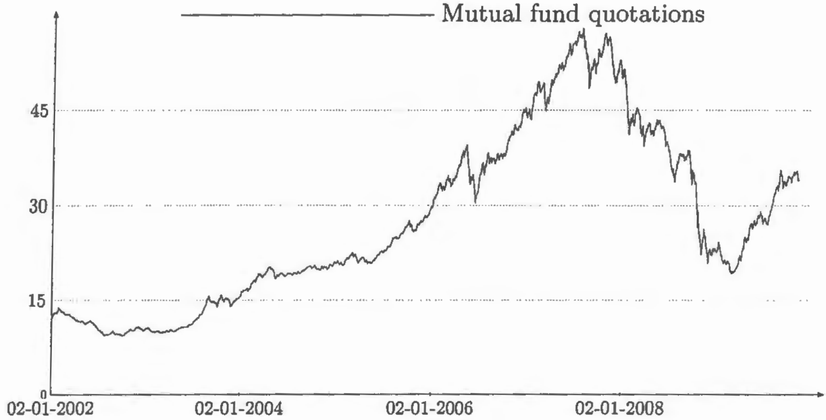

Figure 2: Daily quotations of an investment fund in question

33.29 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.

We obtain 353 extracted trends, with the shortest of 1 time unit only the longest – 71. We assume 3 labels only for each attribute.

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 $\mathcal{I} = 0.1166$, and "Among all decreasing y's, most are short and low", with $\mathcal{I} = 0.1333$, the latter, although it has a bit

Table 1: Some results obtained for 3 labels only for each attribute

linguistic summary	$\mathcal{T}$	d_f	$\mathcal{I}$
Among all low y's, most are short	1.0000	0.7560	0.2736
Among all decreasing y's, almost all are short	1.0000	0.2720	0.1166
Among all increasing y's, almost all are short	1.0000	0.2668	0.1143
Among all short and increasing y's, most are low	1.0000	0.2483	0.1444
Among all decreasing y's, most are low	0.9976	0.2720	0.0596
Among all short and decreasing y's, most are low	0.9969	0.2645	0.1533
Among all increasing y's, most are short and low	0.9860	0.2668	0.1352
Among all y's, most are short	0.9694	–	0.5012
Among all decreasing y's, most are short and low	0.9528	0.2720	0.1333
Among all y's, most are low	0.9121	–	0.3512
Among all short and constant y's, most are low	0.8408	0.2741	0.1597
Among all moderate y's, most are short	0.8274	0.2413	0.0619
Among all constant y's, most are low	0.8116	0.4612	0.1239
Among all medium and constant y's, most are low	0.7646	0.1265	0.0650
Among all medium y's, most are low	0.7167	0.1524	0.0372

smaller truth value, is more informative, as it provides additional information. This is a more general proeprty 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 linguitsic summaries of times series.

5 Concluding remarks

We extended our approach to the linguistic summarization of time series 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. Results on the summarization of quotations of an investment (mutual) fund are very encouraging.

References

[1] New year's eve:past performance is no indication of future return. stockcasting.blogspot.com/2005/12/new-years-evepast-performance-is-no.html.

[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] Ildar 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] Ildar Batyrshin and Leonid Sheremetov. Perception based functions in qualitative forecasting. In Ildar Batyrshin, Janusz Kacprzyk, Leonid Sheremetov, and Lotfi A. Zadeh, editors, *Perception-based Data Mining and Decision Making in Economics and Finance*. Springer-Verlag, Berlin and Heidelberg, 2006.

[6] John C. Bogle. *Common Sense on Mutual Funds: New Imperatives for the Intelligent Investor.* Wiley, New York, 1999.

[7] Janusz Kacprzyk and Anna Wilbik. Linguistic summarization of time series using fuzzy logic with linguistic quantifiers: a truth and specificity based approach. In Leszek Rutkowski, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada, editors, *Artificial Intelligence and Soft Computing ICAISC 2008*, pages 241–252. Springer-Verlag, Berlin and Heidelberg, 2008.

[8] Janusz Kacprzyk and Anna Wilbik. 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*, pages 1146–1153. IEEE Press, 2008.

[9] Janusz Kacprzyk and Anna Wilbik. A new insight into the linguistic summarization of time series via a degree of support: Elimination of infrequent patterns. In D. Dubois, M.A. Lubiano, H. Prade, Maria A. Gil, Przemysaw Grzegorzewski, and Olgierd Hryniewicz, editors, *Soft Methods for Handling Variability and Imprecision*, pages 393–400. Springer-Verlag, Berlin and Heidelberg, 2008.

[10] Janusz Kacprzyk and Anna 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.

[11] Janusz Kacprzyk, Anna Wilbik, and Sł awomir Zadrożny. Linguistic summarization of time series under different granulation of describing features. In Marzena Kryszkiewicz, James F. Peters, Henryk Rybinski, and Andrzej Skowron, editors,

Rough Sets and Intelligent Systems Paradigms - RSEISP 2007, pages 230–240. Springer-Verlag, Berlin and Heidelberg, 2007.

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

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

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

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

[16] Janusz Kacprzyk and Sł awomir Zadrożny. Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries and natural language generation. *IEEE Transactions on Fuzzy Systems.* (forthcoming).

[17] Janusz Kacprzyk and Sł awomir Zadrożny. FQUERY for Access: fuzzy querying for a windows-based dbms. In Patric Bosc and Janusz Kacprzyk, editors, *Fuzziness in Database Management Systems*, pages 415–433. Springer-Verlag, Heidelberg, 1995.

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

[19] Janusz Kacprzyk and Sł awomir 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] Eammon Keogh, Selina Chu, David Hart, and Michael Pazzani. An online algorithm for segmenting time series. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, 2001.

[21] Eammon Keogh, Selina Chu, David Hart, and Michael 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] Randy 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] U.S. Securities and Exchange Commission. Mutual fund investing: Look at more than a fund's past performance. www.sec.gov/investor/pubs/mfperform.htm.

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

[25] S. Sripada, E. Reiter, and I. Davy. Sumtime-mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10, 2003.

[26] Ronald R. Yager. Measuring tranquility and anxiety in decision making: An application of fuzzy sets. *International Journal of General Systems*, 8:139–146, 1982.

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

[28] Ronald R. Yager. On linguistic summaries in data. In Gregory Piatetsky-Shapiro and William J. Frawley, editors, *Knowledge Discovery in Databases*, pages 347–363. MIT Press, Cambridge, USA, 1991.

[29] Ronald R. Yager. On measures of specificity. In Okyay Kaynak, Lotfi A. Zadeh, Burhan Trksen, and Imre J. Rudas, editors, *Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Applications*, pages 94–113. Springer-Verlag: Berlin, 1998.

[30] Ronald R. Yager, Kenneth M. Ford, and Alberto J. Cañas. An approach to the linguistic summarization of data. pages 456–468. Springer, 1990.

[31] Lotfi 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.

[32] Lotfi 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.

[33] Lotfi A. Zadeh. Computation with imprecise probabilities. In *In IPMU'08, Torremolinos (Malaga), June 22-27, 2008*, 2008.