Using fuzzy linguistic summaries for the comparison of time series: an application to the analysis of investment fund quotations

Janusz Kacprzyk[1,2] Anna Wilbik[1*]

[1] Systems Research Institute, Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
[2] WIT – Warsaw School of Information and Technology
Newelska 6, 01-447 Warsaw, Poland
Email: kacprzyk@ibspan.waw.pl, wilbik@ibspan.waw.pl

Abstract— *We propose a new, human consistent method for the evaluation of similarity of time series that uses a fuzzy quantifier base aggregation of trends (segments), within the authors' (cf. Kacprzyk, Wilbik, Zadrożny [1, 2, 3, 4, 5, 6] or Kacprzyk, Wilbik [7, 8, 9]) approach to the linguistic summarization of trends based on Zadeh's protoforms and fuzzy logic with linguistic quantifiers. The results obtain are very intuitively appealing and justified by valuable outcomes of similarity analyses between quotations of an investment fund and the two main indexes of the Warsaw Stock Exchange.*

1 Introduction

Among various application areas of advanced data mining and knowledge discovery tools and techniques, all kinds of finance related areas are of primordial and growing importance. For instance, in an interesting statistical report [10] 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).

This trend will presumably continue in view of serious difficulties on financial/investment market which is expected to continue well after 2009.

This paper is a continuation of our previous works (cf. Kacprzyk, Wilbik, Zadrożny [1, 2, 3, 4, 5, 6] or Kacprzyk, Wilbik [7, 8, 9]) which deal with the problem of how to effectively and efficiently support a human decision maker in making decisions concerning investments. We deal mainly with investments in investment (mutual) funds. Clearly, decision makers are here concerned with possible future gains/losses, and their decisions is related to what might happen in the future. However, our aim is not the forecasting of the future daily prices, which could have been eventually used directly for a purchasing decision. Instead, in our works, we follow a decision support paradigm, that is we try to provide the decision maker with some information that can be useful for his/her decision on whether and how many units to purchase. We do not intend to replace the human decision maker.

This problem is very complex. First of all, there may be two general approaches. The first one, which may seem to be the most natural is to provide means to derive a price forecast for an investment unit so that the decision maker could "automatically' purchase what has been forecast, and as much as he/she could afford. Unfortunately, the success in such a straightforward approach is much less than expected. Basically, statistical methods employed usually for this purpose are primitive in the sense that they just extrapolate the past and do not use domain knowledge, intuition, some inside information, etc. A natural solution may be to try to just support the human decision maker in making those investment decisions by providing him/her with some additional useful information, and not getting involved in the actual investment decision making.

Various philosophies in this respect are possible. Basically, from our perspective, the following one 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 clearly indicates 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.

It should be noted that 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" [11], 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" [12], 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

*Partially supported by the Ministry of Science and Higher Education Grant no. NN 516 4309 33.

ISBN: 978-989-95079-6-8

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 apparently is, 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" [13], 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 [14] 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 ...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: Bogle [15] 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 [16], 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" [17], they say in the last, tenth, 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".

One can give many more quotations from all kinds of investment guides, analyses, etc. by leading experts. Virtually all of them emphasize the importance of looking at the past to help make future decisions. Moreover, they also generally advocate a more comprehensive look in the sense that what might be useful would rather be not particular single values but some deeper meaning, even nature of past behavior and returns.

We have followed this line of reasoning in our past papers (cf. Kacprzyk, Wilbik, Zadrożny [1, 2, 3, 4, 5, 6] or Kacprzyk, Wilbik [7, 8, 9]), 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.

However, for many more detailed analyses we need to relate the performance of an investment fund to the performance (time evolution) of some benchmark(s) against which the particular investment fund is to be compared, as stated in its prospectus. These benchmarks are usually some stock exchange indexes, or a composition thereof.

One of more interesting and relevant questions that may be posed by both a professional analyst or a customer whose money is being invested may be: how similar was the temporal evolution of quotations of the particular investment fund and its benchmark(s)? This has clearly to do with the measuring of similarity of time series.

The problem of similarity of time series or finding similar subsequences of time series is an important issue in many problems, e.g., in indexing [18, 19], clustering [20] or motif discovery [21, 22]. A similarity measure should allow for the imprecise match, i.e., should indicate time series sequences that are similar to some extent. Moreover, the algorithm should be very efficient [23], as we may consider very long time series.

Different approaches to the similarity measures of time series were considered and proposed in the literature. A simple solution is to view each sequence as a point in an n-dimensional Euclidean space, where n is the length of the segment (trend). The similarity or dissimilarity is based on the Euclidean distance, and computed as the norm in this space [24]. This approach can be used for various time series but normally some extensions should be applied providing various transformations such as scaling or shifting were proposed [21].

Another methods use, for instance, the Dynamic Time Wraping (DTW) [19] and Longest Common Subsequence (LCS) measures [24], and they allows gaps in sequences. They are successfully used in other domains such as speech recognition and text pattern matching [25].

Other approaches developed for similarity search in time series data are based on dimension reduction. This can be achieved in numerous ways, e.g. using the Discrete Fourier Transform (DFT) [23], wavelets [26], etc.

One can also mention here approaches based on the piecewise linear representation of time series [27]. In this approach time series are represented by segments (trends), not the particular values. The method proposed here falls into this category.

In this paper we propose new method of calculating the similarity of time series; as we have already indicated, this concerns the quotations of an investment fund and some benchmark. The degree of similarity of two time series is meant here as the degree to which, for instance, "most" of the long, overlapping segments are similar (for instance, such that "almost all" of their features are similar). Notice that this has very much to do with a soft definition of consensus meant as a degree to which, for instance, "*most* of the *important* individuals agree as to *almost all* of the *relevant* options". This was introduced by Kacprzyk and Fedrizzi [28, 29, 30], Kacprzyk, Fedrizzi and Nurmi [31], and recently by Kacprzyk and Zadrożny [32] and has been found useful in many areas.

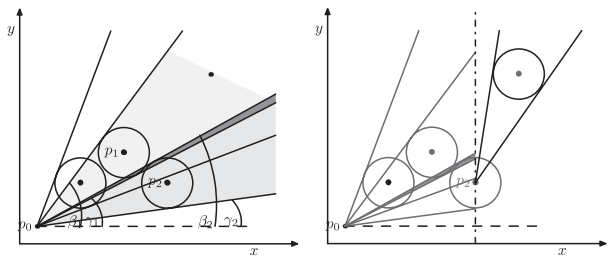
We will now discuss first the segmentation of time series, as the similarity will proceed in terms of segments (trends), not all the consecutive values. Then, we will propose some new, linguistic quantifier based measures of similarity, followed by examples on real data, and conclusions.

ISBN: 978-989-95079-6-8

2 Segmentation of time series

In our approach (cf. Kacprzyk, Wilbik, Zadrożny [1, 2, 3, 4, 5, 6] or Kacprzyk, Wilbik [7, 8, 9]) 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 a 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 [33, 34]). We may either divide the time series into k segments (using, e.g., bottom-up or top-down strategies), where k is specified by the user, or divide it so that the maximum error for any segment does not exceed some threshold value ε, also specified by the user. In our works [1, 2, 3, 4, 5, 6, 7, 8, 9] we used a simple on-line algorithm, a modification of one proposed by Sklansky and Gonzalez [35] in the contest of image analysis. The method is simple and fast, and fully satisfies our needs though many other methods are possible.

Our algorithm constructs the segments so that the maximum distance between the segment line and the observed time point is smaller than a user specified ε value. It works by constructing the intersection of cones starting from point p_i of the time series and including a circle of radius ε around the subsequent data points p_{i+j}, $j = 1, 2, \ldots$, until the intersection of all cones starting at p_i is empty. If for p_{i+k} the intersection is empty, then we construct a new cone starting at p_{i+k-1}. Figure 1 present the idea of the algorithm. The family of possible solutions is indicated as a gray area. Clearly other algorithms can also be used, and there is a lot of them in the literature (cg. [18, 36]).



(a) the intersection of the cones is indicated by the dark grey area

(b) a new cone starts in point p_2

Figure 1: Illustration of the segmentation algorithm used

To present details of the algorithm, let us first denote:
- `p_0` – a point starting the current cone,
- `p_1` – the last point checked in the current cone,
- `p_2` – the next point to be checked,
- `Alpha_01` – a pair of angles (γ_1, β_1), meant as an interval, that defines the current cone as shown in Figure 1(a),
- `Alpha_02` – a pair of angles of the cone starting at the point `p_0` and inscribing the circle of radius ε around the point `p_2` (cf. (γ_2, β_2) in Figure 1(a)),
- function `read_point()` reads a next point of data series,
- function `find()` finds a pair of angles of the cone starting at the point `p_0` and inscribing the circle of radius ε around the point `p_2`.

A pseudocode of the algorithm that extracts trends is depicted in Figure 2.

```
read_point(p_0);
read_point(p_1);
while(1)
{
  p_2=p_1;
  Alpha_02=find();
  Alpha_01=Alpha_02;
  do
  {
    Alpha_01 = Alpha_01 ∩ Alpha_02;
    p_1=p_2;
    read_point(p_2);
    Alpha_02=find();
  } while(Alpha_01 ∩ Alpha_02 ≠ ∅);
  save_found_trend();
  p_0=p_1;
  p_1=p_2;
}
```

Figure 2: Pseudocode of the modified Sklansky and Gonzalez [35] algorithm for extracting trends

The bounding values of `Alpha_02` (γ_2, β_2), computed by function `find()` correspond to the slopes of two lines that (1) are tangent to the circle of radius ε around point $p_2 = (x_2, y_2)$ and (2) start at the point $p_0 = (x_0, y_0)$. Thus

$$\beta_2, \gamma_2 = arctg\left(\frac{\Delta x \cdot \Delta y - \pm\varepsilon\sqrt{(\Delta x)^2 + (\Delta y)^2 - \varepsilon^2}}{(\Delta x)^2 - \varepsilon^2}\right)$$

where $\Delta x = x_0 - x_2$ and $\Delta y = y_0 - y_2$.

The resulting linear ε-approximation of a group of points `p_0, ..., p_1` is either a single segment, chosen as, e.g., a bisector of the cone, or one that minimizes the distance (e.g., the sum of squared errors, SSE) from the approximated points, or the whole family of possible solutions, i.e., rays of the cone.

This method is effective and efficient as it requires only a single pass through the data. From now on we will identify (partial) *trends* with the line segments of the constructed piecewise linear function.

3 Linguistic summarization of time series

First, we will outline the very essence of our approach to the linguistic summarization of time series which is based on the two types of protoforms of linguistic summaries of trends:
- a short form:

$$\text{Among all segments, } Q \text{ are } P \tag{1}$$

e.g.: "Among all segments, *most* are *slowly increasing*".
- an extended form:

$$\text{Among all } R \text{ segments, } Q \text{ are } P \tag{2}$$

e.g.: "Among all *short* segments, *most* are *increasing*".
whose truth values are, respectively:

$$\mathcal{T}(\text{Among all } y\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_P(y_i)\right) \tag{3}$$

$$\mathcal{T}(\text{Among all } Ry\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_R(y_i) \wedge \mu_P(y_i)}{\sum_{i=1}^{n}\mu_R(y_i)}\right) \tag{4}$$

where $\wedge$ is the minimum operation (more generally it can be another appropriate operator, notably a t-norm).

ISBN: 978-989-95079-6-8

For more information on this approach, and on the advantages of using protoforms as general forms of linguistic summaries, see Kacprzyk, Wilbik, Zadrożny [1, 2, 3, 4, 5, 6] or Kacprzyk, Wilbik [7, 8, 9].

4 A new fuzzy quantifier based method for the evaluation of similarity of time series

The degree of similarity of two time series is meant here as the degree to which, say "most" of the long, simultaneous segments are similar (i.e., "majority" of their features are similar).

The time series is represented by its segments (trends). Each segment is described by the following four values:
• starting time, called here a partition point,
• duration of the segment (expressed in time units),
• dynamics of change (expressed as the slope of the segment),
• variability occurring in the segment, describing the variability of differences between the trend (segment line) and the particular time series values over the duration of this trend.

Let us assume, we wish to compare two time series A and B. As the result of the segmentation procedure time series A was divided into k segments, represented by a_j, $j = 1, \ldots, k$ and time series B was divided into l segments, represented by b_j, $j = 1, \ldots, l$.

Next we create the set of partition points of the two time series considered, i.e. this set contains the time points when at least one segment has started. There are at most $k + l$ such partition points. Those times are sorted, so that the earlier ones come first.

Now between any consecutive partition points (e.g. p_i and p_{i+1}) there is only one segment (or a part of it) of each time series. We may denote those segments as a_{p_i} and b_{p_i}, i.e., a_{p_i} is the segment representing time series A that take place between p_i and p_{i+1}. Now we can compute the similarity of such simultaneous segments.

We compute the degree of similarity of two segments (trends), as the degree to which "most" of their features are similar. Here we consider for simplicity the three features only: duration, dynamics and variability. The values of the features of segment a_{p_i} are $dur_{a_{p_i}}$ for duration, $dyn_{a_{p_i}}$ for dynamics, and $var_{a_{p_i}}$ for variability. R_{dur}, R_{dyn} and R_{var} are the ranges of possible values of duration, dynamics and variability. The degree of similarity of segments a_{p_i} and b_{p_i} with respect to duration is computed as

$$sim^{dur}(a_{p_i}, b_{p_i}) = \mu^{dur}_{sim}\left(\frac{|dur_{a_{p_i}} - dur_{b_{p_i}}|}{R_{dur}}\right) \quad (5)$$

Similarly, separate functions can be defined for the similarity for each feature, here μ^{dyn}_{sim} and μ^{var}_{sim}. Let $\mu_{Q_1}(.)$ be the membership function of quantifier "most", i.e regular, nondecreasing and monotone. Thus the degree of similarity of segments a_{p_i} and b_{p_i} is computed as

$$\begin{aligned} sim(a_{p_i}, b_{p_i}) = \mu_{Q_1}\Bigg(& w_{dur}\left(\mu^{dur}_{sim}\left(\frac{|dur_{a_{p_i}} - dur_{b_{p_i}}|}{R_{dur}}\right)\right) + \\ &+ w_{dyn}\left(\mu^{dyn}_{sim}\left(\frac{|dyn_{a_{p_i}} - dyn_{b_{p_i}}|}{R_{dyn}}\right)\right) + \\ &+ w_{var}\left(\mu^{var}_{sim}\left(\frac{|var_{a_{p_i}} - var_{b_{p_i}}|}{R_{var}}\right)\right)\Bigg) \end{aligned} \quad (6)$$

where $w_{dur} + w_{dyn} + w_{var} = 1$, here all are equal, $w_{dur} = w_{dyn} = w_{var} = \frac{1}{3}$.

Next, to obtain the similarity of A and B, we need to aggregate above similarity values between the overlapping segments, and again we use a linguistic quantifier driven aggregation.

$$sim(A, B) = \mu_Q\left(\frac{\sum_{i=1}^{n} \frac{p_{i+1}-p_i}{T} sim(a_{p_i}, b_{p_i})}{\sum_{i=1}^{n} \frac{p_{i+1}-p_i}{T}}\right) \quad (7)$$

where $p_{i+1} - p_i$ is the difference between the two consecutive partition points p_i and p_{i+1}, and T is the total time span considered.

As $\sum_{i=1}^{n} p_{i+1} - p_i = T$ we may simplify this formula and we obtain

$$sim(A, B) = \mu_Q\left(\sum_{i=1}^{n} \frac{p_{i+1} - p_i}{T} sim(a_{p_i}, b_{p_i})\right) \quad (8)$$

5 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 (WSE). We have used two benchmark time series, the index of WSE companies (WIG), that is the benchmark for this fund mentioned in the fund prospectus, and the index of 20 biggest and most liquid companies (WIG20). More information can be found on the WSE Web page (`http://www.gpw.pl`).

Data shown in Fig. 3 were collected from the beginning of 2007 until the end of 2008 (501 quotations); notice that one can see when world wide financial problems begun. They are measured in different units, and in order to compare them, we have scaled them using the following formula. The new value of the quotation

$$\bar{v}_i = \frac{v_i - v_0}{v_0} 10 + 10$$

where v_0 is the quotation value of the first observation, i.e. on January, 2, 2007. Those scaled time series may be interpreted as the amount of money that would be earned if we invested PLN 10 (PLN stands for the Polish Zloty) in the mutual fund or stocks of the index on January 2, 2007.

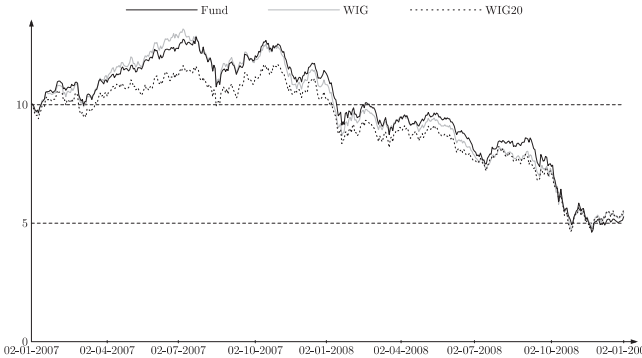


Figure 3: Daily quotations of an investment fund considered and the WIG and WIG20 indexes

The value of the mutual fund time series was equal PLN 5.28 at the end of the time span considered. The minimal value

ISBN: 978-989-95079-6-8

recorded was PLN 4.63 while the maximal one during this period was PLN 12.86.

The values of the benchmark time series were equal to PLN 5.46 and PLN 5.61 at the end of the time span considered for the WIG and WIG20, respectively. The minimum value of WIG recorded was PLN 4.71 while the maximum one during this period was PLN 13.20. The minimum and maximum values for WIG20 were PLN 4.62 and PLN 11.70, respectively.

Using the modified Sklansky and Gonzalez algorithm (cf. [2]) and $\varepsilon = 0.25$ we obtained 43 extracted trends (segments) for the fund time series, 51 for the WIG time series and 49 for the WIG 20 time series.

We can describe each time series independently using linguistic summaries of time series (cf. Kacprzyk, Wilbik, Zadrożny [1, 2, 3, 4, 5, 6] or Kacprzyk, Wilbik [7, 8, 9]). As those data sets serve only the purpose of example and are small, we applied the granulation with 3 labels for each feature only, namely *decreasing*, *constant* and *increasing* for the dynamics of change; *short*, *medium length* and *long* for duration and *low*, *moderate* and *high* for the variability.

The summaries obtained for the fund time series along with their truth values $\mathcal{T}$ are shown in Table 1.

Table 1: Linguistic summaries for the mutual fund time series

linguistic summary	$\mathcal{T}$
Among all y, most are constant	1
Among all medium y, most are constant	1
Among all moderate y, most are constant	1
Among all low y, most are constant	1
Among all medium and moderate y, most are constant	1
Among all medium and low y, most are constant	1
Among all long y, most are constant	1
Among all long and moderate y, most are constant	1
Among all short y, most are low	0.8107
Among all short y, most are constant	0.7819
Among all moderate y, most are medium	0.7462
Among all moderate y, most are medium and constant	0.7462
Among all long y, most are constant and moderate	0.7302
Among all long y, most are moderate	0.7302
Among all long and constant y, most are moderate	0.7302

The summaries obtained for the WIG index time series along with their truth values $\mathcal{T}$ are shown in Table 2.

Table 2: Linguistic summaries for the WIG index time series

linguistic summary	$\mathcal{T}$
Among all y, most are constant	1
Among all medium y, most are constant	1
Among all low y, most are constant	1
Among all moderate y, most are constant	1
Among all medium and low y, most are constant	1
Among all medium and moderate y, most are constant	1
Among all long y, most are constant	1
Among all short and moderate y, most are constant	1
Among all short y, most are constant	0.9950
Among all short and low y, most are constant	0.7856

The summaries obtained for the WIG20 index time series along with their truth values $\mathcal{T}$ are shown in Table 3.

Table 3: Linguistic summaries for the WIG20 index time series

linguistic summary	$\mathcal{T}$
Among all y, most are constant	1
Among all y, most are constant and low	1
Among all y, most are low	1
Among all constant y, most are low	1
Among all low y, most are constant	1
Among all medium y, most are constant	1
Among all medium y, most are constant and low	1
Among all medium y, most are low	1
Among all medium and constant y, most are low	1
Among all medium and low y, most are constant	1
Among all short y, most are low	1
Among all short and constant y, most are low	1
Among all long y, most are constant	1
Among all long y, most are constant and low	1
Among all long y, most are low	1
Among all long and constant y, most are low	1
Among all long and low y, most are constant	1
Among all short and low y, most are constant	0.9713
Among all short y, most are constant	0.9465
Among all short y, most are constant and low	0.8850
Among all constant y, most are medium	0.7333

The degree of similarity between the fund quotation time series and the WIG time series is equal 0.8622, while the degree of similarity of the fund quotation time series and the WIG20 time series is equal to 0.7529. This result clearly confirms our intuition. This may be viewed, on the one hand, as a proof of usefulness of our method, and on the other hand is quite natural because the investment fund considered in based on the WIG index. Luckily enough, the similarity with the WIG20 index, that of the biggest and most liquid companies, is also at a similarly high level.

6 Concluding remarks

We have proposed a new, human consistent method for the evaluation of similarity of time series, within our approach to the linguistic summarizatioon of time series that uses a fuzzy quantifier based aggregation of trends. The results obtained are very intuitively appealing. This is justified by the analysis of similarity between quotations of an investment fund and the two main indexes of the Warsaw Stock Exchange, WIG and WIG20. It also seems that a relation to natural language generation (NLG), along the lines of Kacprzyk and Zadrożny [37], may be very promising.

References

[1] 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.

[2] J. Kacprzyk, A. Wilbik, and S. Zadrożny. On some types of linguistic summaries of time series. In *Proceedings of the 3rd International IEEE Conference "Intelligent Systems"*, pages 373–378. IEEE Press, 2006.

[3] J. Kacprzyk, A. Wilbik, and S. Zadrożny. Linguistic summarization of time series by using the choquet integral. In P. Melin, O. Castillo, L.T. Aguilar, J. Kacprzyk, and W. Pedrycz, editors,

ISBN: 978-989-95079-6-8

Foundations of Fuzzy Logic and Soft Computing - IFSA 2007, pages 284–294. Springer-Verlag, Berlin and Heidelberg, 2007.

[4] 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.

[5] J. Kacprzyk, A. Wilbik, and S. Zadrożny. Mining time series data via linguistic summaries of trends by using a modified sugeno integral based aggregation. In *IEEE Symposium on Computational Intelligence and Data Mining CIDM 2007*, pages 467–472. IEEE Press, 2007.

[6] 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.

[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. 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, M. A. Gil, P. Grzegorzewski, and O. Hryniewicz, editors, *Soft Methods for Handling Variability and Imprecision*, pages 393–400. Springer-Verlag, Berlin and Heidelberg, 2008.

[9] J. Kacprzyk and A. 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.

[10] Subject: Poll results: Top data mining applications in 2008 and trends. http://www.kdnuggets.com/news/2009/n01/1i.html.

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

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

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

[14] 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.

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

[16] 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.

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

[18] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Locally adaptive dimensionality reduction for indexing large time series databases. In *Proceedings of ACM SIGMOD Conference on Management of Data*, pages 151–162, Santa Barbara, CA, 2001.

[19] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge Information Systems*, 7(3):358–386, 2005.

[20] P. Geurts. Pattern extraction for time series classification. In *PKDD '01: Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, London, UK, 2001. Springer-Verlag.

[21] B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *Proceedings of the the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 493–498, 2003.

[22] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover. Exact discovery of time series motifs. In *SDM 2009, in press*, 2009.

[23] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*, pages 69–84, 1993.

[24] G. Das, D. Gunopulos, and H. Mannila. Finding similar time series. In Jan Komorowski and Jan Zytkow, editors, *Principles of Data Mining and Knowledge Discovery (PKDD'97)*, pages 88–100. Springer-Verlag, 1997.

[25] N. Nakatsu, Y. Kambayashi, and S. Yajima. A longest common subsequence algorithm suitable for similar text strings. *Acta Informatica*, 18:171–179, 1982.

[26] K.-P. Chan and W.-C. Fu. Efficient time series matching by wavelets. In *Proceedings of the 15th International Conference on Data Engineering*. IEEE Computer Society, 1999.

[27] H. Shatkay and S. B. Zdonik. Approximate queries and representations for large data sequences. In *Proceedings 12th International Conference on Data Engineering*, pages 536–545, 1996.

[28] J. Kacprzyk and M. Fedrizzi. 'soft' consensus measures for monitoring real consensus reaching processes under fuzzy preferences. *Control and Cybernetics*, 15:309–323, 1986.

[29] J. Kacprzyk and M. Fedrizzi. A 'soft' measure of consensus in the setting of partial (fuzzy) preferences. *European Journal of Operational Research*, 34:315–325, 1988.

[30] J. Kacprzyk and M. Fedrizzi. A 'human-consistent' degree of consensus based on fuzzy logic with linguistic quantifiers. *Mathematical Social Sciences*, 18:275–290, 1989.

[31] J. Kacprzyk, M. Fedrizzi, and H. Nurmi. Group decision making and consensus under fuzzy preferences and fuzzy majority. *Fuzzy Sets and Systems*, 49:21–31, 1992.

[32] J. Kacprzyk and S. Zadrożny. Towards a general and unified characterization of individual and collective choice functions under fuzzy and nonfuzzy preferences and majority via the ordered weighted average operators. *International Journal of Intelligent Systems*, 24(1):4–26, 2009.

[33] 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.

[34] 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.

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

[36] E. Keogh and M. Pazzani. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 239–241, New York, NY, 1998.

[37] 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.

ISBN: 978-989-95079-6-8