LINGUISTIC COMPARISON OF TIME SERIES: A FUZZY APPROACH

Rita Castillo-Ortega Nicolás Marín Daniel Sánchez

Department of Computer Science and A.I.,
University of Granada, 18071, Granada, Spain
{rita,nicm,daniel}@decsai.ugr.es

Abstract

In this paper, we provide a method for the linguistic comparison of time series. The method is based on the linguistic summarization of time series obtained as the difference between the series being compared. Several approaches to compute the difference between the series lead to alternative semantics in the summary. The presented method is applied to an example in the Business Intelligence field.

Keywords: Linguistic Summarization, Time series, Business Intelligence, Fuzzy Logic.

1 INTRODUCTION

Nowadays, it is very important to be able to manage a wide range of information referring to business activities. The process of obtaining information from data is getting more and more important as it allows users to make decision analysis and forecasting [2]. In particular, the research in the Information Systems field has been specially focused on searching for Business Intelligence solutions.

Time dimension usually appears in Information Systems due to its importance in any commercial activity. As an example, dealing with *historical*, current and future predictive views are some of the most important areas within Business Intelligence.

Many authors have focused their research activity on time series Data Mining[1]. As important as the information extraction is the process that permits users a better understanding of this information. In this context, linguistic summarization techniques are of special interest because they produce sentences close to natural language to describe data.

One of the most extended ways to face the linguistic summarization problem is through the use of Fuzzy Set Theory. An essential part of this research can be found in R. R. Yager's works, where he uses quantified sentences in the sense of L. A. Zadeh first [15],[16], and OWA operators later [17]. Also following Zadeh's footsteps, we can find J. Kacprzyk et al. in [7], [11], [10], [12], and [9], proposing new quality measures and using the protoform concept. G. Raschia et. al created a model named SaintEtiQ [13], working with hierarchies. From a different point of view, P. Bosc and D. Dubois proposed the use of association rules [3].

In this area, we have focused our efforts in obtaining linguistic summaries that *briefly* present the *essential* information regarding the evolution of a given feature over time. Our proposal [4, 5] is based on the use of *fuzzy quantified statements* and tries to get a brief summary taking advantage of a fuzzy hierarchical partition of a time dimension.

Another problem in this area where linguistic descriptions are of special interest is the comparison of time series. In [6] M. Umano et al. carry out a study about describing time series data using their global trend and local features. In another related work, J. Kacprzyk and A. Wilbik [8] focus on the evaluation of similarity of time series. They proposed a fuzzy quantifier based aggregation approach and apply their method to the analysis of investment fund quotations in order to show the utility of the technique.

In this paper, a method for the linguistic comparison of time series is presented. The method is based on the application of our linguistic summarization approach to a time series that appropriately describes the difference between the series being compared. We analyze several approaches to compute the difference between the series, leading this way to alternative semantics in the summary.

The paper is organized as follows: Section 2 defines the comparison problem; Section 3 presents the proposed approach to obtain the linguistic comparison. Finally, a practical example appears in Section 4 and some conclusions are presented in Section 5.

2 TIME SERIES COMPARISON ACROSS TIME

The problem we address in this work is, given two time series of data defined over the same time domain, to obtain a linguistic summary that describes the difference in value between the two series across time. As we will see, this linguistic summary is achieved by means of a hierarchical fuzzy partition of the time domain and a fuzzy linguistic granulation of the variable domain. An illustrative example of these elements is depicted in Figure 1.

As we can see, the figure represents the behavior of a given variable V along *time* in two different series. The y-coordinate displays the domain of the variable V as a fuzzy partition. The x-coordinate shows the time dimension hierarchically organized as a set of different fuzzy partitioned levels.

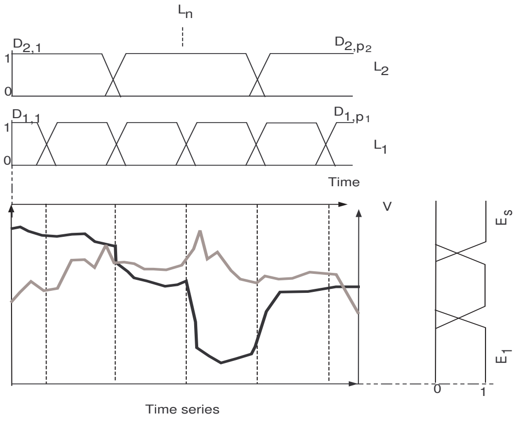

Figure 1: Time series

We assume that the time dimension is described in its finest grained level of granularity by $T = \{t_1, ..., t_m\}$. Then, we consider a couple of time series defined on this time dimension, namely, TS_1 and TS_2. $TS_i(t_j)$ represents the value of the variable V under study in t_j as given by TS_i.

As we have previously mentioned:

- The basic domain of variable V is partitioned by a set of linguistic labels E=$\{E_1, ..., E_s\}$.
- The time dimension is hierarchically organized in n levels, namely, L=$L_1, ..., L_n$. Each level L_i has associated a partition $\{D_{i,1}, ..., D_{i,p_i}\}$ of the basic time domain.

There is no restriction concerning the form of the membership function of a label apart from that it must be normalized. In our approach, we will use trapezoidal functions. When necessary, labels $D_{i,j}$ in time dimension can be the union of a set of trapezoidal functions.

In this work, a set of labels $\{X_1, ..., X_r\}$ is a partition on X iff:

1. $\forall x \in X, \exists X_i, i \in \{1..r\} | \mu_{X_i}(x) > 0$.
2. $\forall i, j \in \{1..r\}, i \neq j, core(X_i) \cap core(X_j) = \emptyset$.

Additionally, concerning the hierarchy of the time dimension, we add the following constraints:

1. $\forall i, j \in \{1..n\}, i < j, p_i > p_j$ (i.e, as we move upward in the hierarchy, the number of labels of the partition decreases).
2. $\forall i \in \{2..n\}, \forall j \in \{1..p_i\}, \forall k \in \{1..p_{i-1}\} | (D_{i,j} \subseteq D_{i-1,k}) \rightarrow (D_{i,j} = D_{i-1,k})$ (i.e., labels cannot generalize another label of an upper level).

3 THE PROPOSED METHOD

In order to face the problem of series comparison, we propose the obtention of a new time series as the difference ΔTS between the two series being compared. Then we can linguistically summarize this new series obtaining information regarding the comparison. These two steps are detailed below.

3.1 Obtaining the ΔTS time series

ΔTS time series can be obtained in several different ways. Let us consider the following definition.

Definition 1 *Let TS_1 and TS_2 be two time series defined over the same variable V at a given period of time. Then,*

$$\Delta TS_{abs}(t_i) = TS_1(t_i) - TS_2(t_i) \quad (1)$$

$$\Delta TS_{global}(t_i) = \begin{cases} 0, if\ TS_1(t_i) - TS_2(t_i) = 0 \\ \dfrac{TS_1(t_i) - TS_2(t_i)}{M - m}, otherwise \end{cases} \quad (2)$$

$$\Delta TS_{local}(t_i) = \begin{cases} 0, if\ TS_1(t_i) - TS_2(t_i) = 0 \\ \dfrac{TS_1(t_i) - TS_2(t_i)}{max(TS_1(t_i), TS_2(t_i)) - m}, otherwise \end{cases} \quad (3)$$

where t_i is a specific point in the time domain, M is the global maximum of TS_1 and TS_2, and m is the global minimum of TS_1 and TS_2.

This definition proposes three different ways to face the computation of the new series that describes the difference between the two original ones. At a given time point t_i,

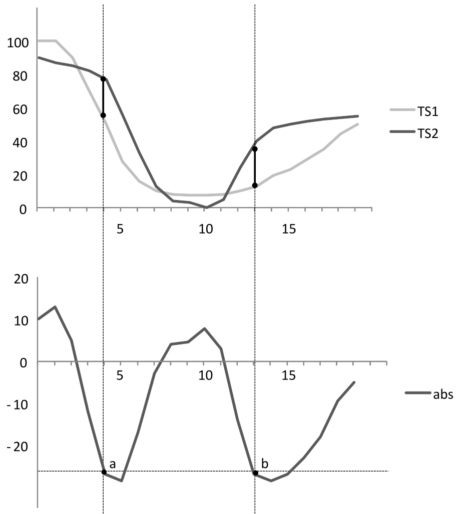

Figure 2: Time series data TS_1, TS_2 and ΔTS_{abs}

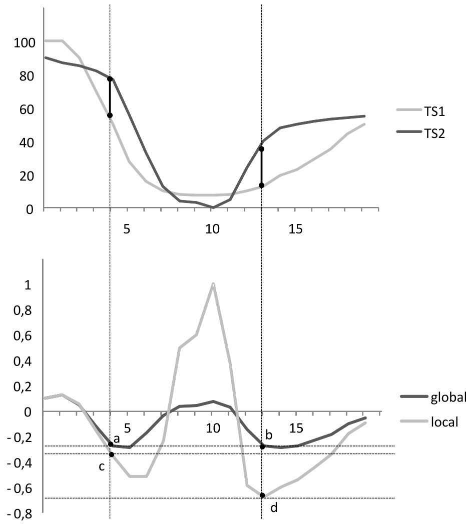

Figure 3: ΔTS_{local} and ΔTS_{global}

- $\Delta TS_{abs}(t_i)$ is the difference, in absolute terms, between the two original series at the point t_i (Equation 1).
- $\Delta TS_{global}(t_i)$ is the difference, in relative terms, between the two original series at the point t_i, according to the scale of values of the two original series (i.e., the difference between the global maximum and minimum of the two series)(Equation 2).
- $\Delta TS_{local}(t_i)$ is the difference, also in relative terms, between the two original series at the point t_i, but now according to the scale of values of the given time point in the two original series (i.e., the difference between the maximum value at the given time point and the global minimum) (Equation 3).

The choice of the strategy depends on the necessities of the user or the problem in each particular situation. ΔTS_{abs} is the only option if we are interested in the analysis of the difference between the series in absolute terms. Figure 2 depicts an example of the use of this first alternative. As can be seen, the new series ranges over the same variable domain than the original one.

Nevertheless, if we are interested in the analysis of the difference between the series in relative terms, we have two different alternatives. Figure 3 shows the ΔTS_{global} and ΔTS_{local} obtained for the previous example. The figure illustrates the difference between the two strategies: while in ΔTS_{global} the same difference between the original series produces always the same value in the new one (see points a and b), in ΔTS_{local} the *lower* the original values the greater the *significance* of the difference (see points c and d). In this sense, ΔTS_{abs} behaves like ΔTS_{global} (see Figure 2, points a and b).

Additionally, in contrast to the case of ΔTS_{abs}, the new series (both ΔTS_{global} and ΔTS_{local}) range over $[-1, 1]$.

Once we have opted for a given strategy, we have to provide a linguistic description of the domain of the new series. In our method, we use a linguistic variable with the following features:

- The linguistic variable covers both positive and negatives values.
- In the case of ΔTS_{abs}, it is defined on the range $[-M, M]$.
- In the case of ΔTS_{global} and ΔTS_{local}, it is defined on the range $[-1, 1]$.

As an example, Figure 4 depicts a possible linguistic variable for the interval $[-1, 1]$.

3.2 Obtaining the summary

In the previous subsection, we have analyzed how to obtain a new time series that describes the difference

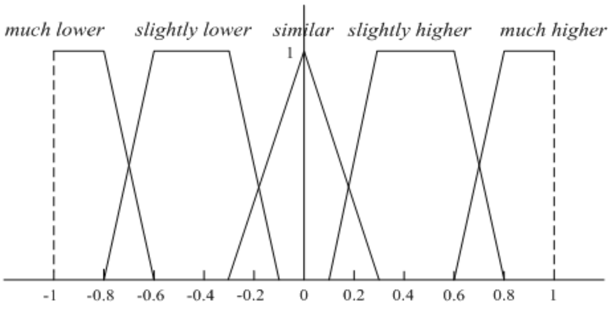

Figure 4: Linguistic labels.

between the two time series under study. Now, we have to obtain an appropriate linguistic summary of this new series.

We consider that the user is interested in linguistic summaries that take the form of a collection of quantified sentences describing the behavior of a time series. We assume that the basic elements of these summaries are the linguistic labels described in Section 2. That is, our approach will deliver a collection of sentences of the form "Q of $D_{i,j}$ are A" where:

- $D_{i,j}$ is a label member of a certain level i of the hierarchy associated to the time dimension.
- A is a label or the union of a subset of labels of the partition of the variable V under study (in our case, according to the strategy followed in order to obtain the new series, this partition will be defined on [-M,+M] or [-1,1]).

The user must provide a collection of quantifiers defining the kind of fuzzy quantities and percentages she/he is interested in. This can be defined by choosing among a collection of predefined quantifiers. In this work, we consider that the user provides a *totally ordered* subset $\{Q_1, ..., Q_{qmax}\}$ of a coherent family of quantifiers Q [14] to be used in the summarization process. In addition, the user will provide a threshold τ for the minimum accomplishment degree he wishes for the quantified sentences comprising the summaries.

Figure 5 represents one of the algorithms we proposed in [5] for obtaining the linguistic summaries. In order to look for brevity of the summary, we start from the time periods in the top level of the hierarchy. Each level has its own quantifier bound ($Qbound_i$) and grouping bound ($Gbound_i$) that, respectively, indicate the less strict quantifier to be considered and the maximum number of labels E_i to be aggregated in a sentence at this level of the time domain. The set $ToSummarize$ is the collection of time periods for which a quantified sentence is missing. If it is possible to obtain an accomplishment degree greater than τ for a certain period using a quantifier Q and a single label, the procedure obtains a summary for that period. If it is not possible, the procedure tries with the union of different subsets of labels: couples, trios, quartets, etc, until we obtain an accomplishment degree greater than τ. The size of the subset is given by k being $Gbound$ its maximum value. When a summary is found in a certain time period we say that the period is *covered*. If all the groups were tried without success, the algorithm repeats the grouping process again, but with a less strict quantifier, until $Qbound_i$ is reached. If no result is found for a given period $D_{i,j}$, we try to obtain such sentences with the *corresponding children* $ch(D_{i,j})$ in the next level. If $ch(D_{i,j}) = \emptyset$, then a sentence indicating the observed variability is added to the summary ($D_{i,j}$ *is highly variable* [1]). The final set of linguistically quantified sentences comprising the summary is $Summary$.

```
1. $ToSummarize \leftarrow L_n$;
2. $Summary \leftarrow \emptyset$; $Summarized \leftarrow \emptyset$;
3. While $ToSummarize \neq \emptyset$
   (a) Take $D_{i,j} \in ToSummarize$
   (b) $ToSummarize \leftarrow ToSummarize \backslash \{D_{i,j}\}$
   (c) $p \leftarrow qmax$; $covered \leftarrow false$;
   (d) While $p \geq Qbound_i$ and $not\ covered$
       i. $k \leftarrow 1$;
       ii. While $k \leq Gbound_i$ and $not\ covered$
           A. Let $A \leftarrow argmax_{B \in C_k} GD_{Q_p}(B/D_{i,j})$
           B. If $GD_{Q_p}(B/D_{i,j}) \geq \tau$ then
                 $Summary \leftarrow Summary \cup \{Q_p$ of $D_{i,j}$ are $A\}$;
                 $Summarized \leftarrow Summarized \cup (D_{i,j})$
                 $covered \leftarrow true$;
           C. $k \leftarrow k + 1$;
       iii. $p \leftarrow p - 1$
   (e) If $not\ covered$ and $i > 1$ then
       $ToSummarize \leftarrow ToSummarize \cup ch(D_{i,j})$.
   (f) else if $i = 1$ then
       $Summary \leftarrow Summary \cup \{D_{i,j}$ is highly variable$\}$
```

where $ch(D_{i,j})$ is defined as follows: $ch(D_{1,j}) = \emptyset$ for all j. Otherwise, $ch(D_{i,j}) = \{D_{i-1,k}, k \in \{1..p_{i-1}\} | D_{i-1,k} \cap D_{i,j} \neq \emptyset$ and $\neg \exists D \in ToSummarize \cup Summarized$, $(D_{i-1,k} \cap D_{i,j}) \subseteq D\}$, and $C_k = \{\cup_{E_h \in F} E_h \mid F \subseteq E, |F| = k\}$.

Figure 5: Algorithm 1 to obtain linguistic summaries.

The second algorithm proposed in [5] is similar to that in Figure 5 with the difference that when it is not possible to obtain an accomplishment degree greater than τ for a certain period using a quantifier Q and a single label, the procedure tries with less strict quantifiers before trying with aggregations of labels in E. In

[1] Though variability is also analyzed in [9], we refer to this characteristic of the series only when we cannot suitably summarize the values for a given time period.

summary, the first approach tries to obtain summaries with more restrictive quantifiers, while the second one tries to obtain summaries with smaller groups of labels. The obtained results will depend on the family of quantifiers, the threshold τ, the bounds pointed by the user at the different levels of the time hierarchy, and, of course, the data series.

When the summary is obtained, the set of sentences is postprocessed in order to produce an easier to read paragraph. The repetitions can be removed via merging sentences that cover different time periods but produce the same trend. The merging process takes into account sentences with the same quantifier and trend as well as sentences with the same quantifier. For example, if the summary is *In A, the difference was mostly low or very low. In B, the difference was mostly low or very low*, the function produces the sentence *In A and B, the difference was mostly low or very low*, which is shorter and clearer.

4 A PRACTICAL CASE

In this section we will work with a representative example in order to apply our process and clarify concepts or procedures introduced in formers sections. Figure 6 contains the time series that shows the patient inflow on two different medical centers during a given year.

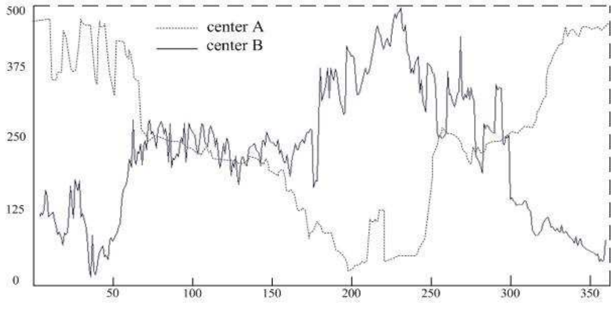

Figure 6: Patient inflow in two medical centers.

Using equations 2 and 3 we obtain the time series data depicted in Figure 7, both relative differences. The y-coordinate represents the differences in terms of Figure 4, and the x-coordinate represents the time dimension. As we can see, the time dimension is hierarchically organized following a meteorological criteria thanks to three fuzzy partition of the time domain, namely: one based on approximate months (in order to avoid a strong dependence of the obtained summaries with respect to the crisp boundaries of conventional months) and two others with different levels of granularity. Fuzziness is specially useful in these two last partitions because transitions between periods are clearly fuzzy.

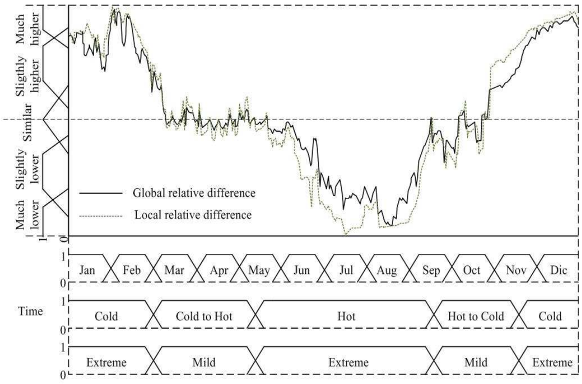

Figure 7: Relative differences.

The example is carried out using a subset of trapezoidal quantifiers $Q = \{Q_1 = (0, 0.4, 0.6, 1),\ Q_2 = (0, 0.6, 0.8, 1),\ Q_3 = (0, 0.7, 0.9, 1)\}$ (that we have called *at least half of*, *at least 70% of*, and *most of*, respectively) and a threshold $\tau = 0.7$ and values $Qbound_i = Gbound_i = 2\ \forall i$ as parameters.

For instance, we will use algorithm 1 (Figure 5) with the time data taken from the *local* relative difference (ΔTS_{local}) between patient inflow in centers A and B. The set of quantified sentences obtained is:

- At least 70% of the time, the patient inflow with cold weather is much higher in center A than in center B
- At least 70% of the time, the patient inflow with hot weather is slightly lower or much lower in center A than in center B
- At least 70% of the time, the patient inflow in March is slightly higher or similar in center A than in center B
- The patient inflow difference between center A and center B in April presents variability
- The patient inflow difference between center A and center B in May presents variability
- The patient inflow difference between center A and center B in September presents variability
- At least 70% of the time, the patient inflow in October is similar or slightly lower in center A than in center B
- Most of the time, the patient inflow in November is much higher or slightly higher in center A than in center B

After the postprocessing, we will have

Most of the time, the patient inflow in November is much higher or slightly higher in center A than center B. At least 70% of the time, the patient inflow with cold weather is much higher in center A than center B, with hot weather is slightly lower or much lower in center A than center B, and in March is slightly higher or similar in center A than center B. Finally, the patient inflow difference in April, *May and September presents variability.*

5 CONCLUSIONS

We have presented an approach to obtain a particular case of linguistic comparison of two time series, as a summary of the difference in each point between both series. This technique is relevant as a Business Intelligence tool for decision making, since it provides an easy-to-understand information to the decision maker. It is worth noticing that understandability and brevity are improved by using user-defined hierarchical fuzzy partition of time dimension and user-defined quantifiers. As future work we plan to perform linguistic comparison on the basis of other features, and to apply new summarization algorithms that are currently being developed.

Acknowledgements

The research reported in this paper was partially supported by the Andalusian Government (Junta de Andalucía) under project P07-TIC03175 and also by the Spanish Government (Science and Innovation Department) under project TIN2009-08296.

References

[1] I.Z. Batyrshin and L. Sheremetov. Perception-based approach to time series data mining. *Appl. Soft Comput.*, 8(3):1211–1221, 2008.

[2] I.Z. Batyrshin and T. Sudkamp. Perception based data mining and decision support systems. *International Journal of Approximate Reasoning*, 48(1):1–3, 2008.

[3] P. Bosc, D. Dubois, O. Pivert, H. Prade, and M. De Calmes. Fuzzy summarization of data using fuzzy cardinalities. In *Int. Conf. Inf. Process. Manag. Uncertainty Knowl. Based Syst.*, pages 1553–1559, 2002.

[4] R. Castillo-Ortega, N. Marín, and D. Sánchez. Fuzzy quantification-based linguistic summaries in data cubes with hierarchical fuzzy partition of time dimension. In H. Yin and E. Corchado, editors, *IDEAL'09*, volume 5788 of *LNCS*, pages 578–585. Springer, Heidelberg, 2009.

[5] R. Castillo-Ortega, N. Marín, and D. Sánchez. Linguistic summary-based query answering on data cubes with time dimension. In T. Andreasen et al., editor, *FQAS'09*, volume 5822 of *LNAI*, pages 560–571. Springer, Heidelberg, 2009.

[6] M. Humano, M. Okamura, and K. Seta. Improved method for linguistic expression of time series with global trend and local features. In *FUZZ-IEEE 2009*, pages 1169–1174, 2009.

[7] J. Kacprzyk. Fuzzy logic for linguistic summarization of databases. In *IEEE International Fuzzy Systems Conference*, pages 813–818, 1999.

[8] J. Kacprzyk and A. Wilbik. Using fuzzy linguistic summaries for the comparison of time series: an application to the analysis of investment fund quotations. In U. Kaymak J. P. Carvalho, D. Dubois and J. M. C. Sousa, editors, *IFSA-EUSFLAT 2009*, pages 1321–1326, 2009.

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

[10] J. Kacprzyk and R. R. Yager. Linguistic summaries of data using fuzzy logic. In *International Journal of General Systems*, volume 30, pages 133–154, 2001.

[11] J. Kacprzyk, R. R. Yager, and S. Zadrozny. A fuzzy logic based approach to linguistic summaries in databases. *International Journal of Applied Mathematical Computer Science*, 10:813–834, 2000.

[12] J. Kacprzyk and S. Zadrozny. Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Inf. Sci. Inf. Comput. Sci.*, 173(4):281–304, 2005.

[13] G. Raschia and N. Mouaddib. Saintetiq: a fuzzy set-based approach to database summarization. *Fuzzy Sets Syst.*, 129(2):137–162, 2002.

[14] M. A. Vila, J. C. Cubero, J. M. Medina, and O. Pons. The generalized selection: an alternative way for the quotient operations in fuzzy relational databases. In B. Bouchon-Meunier, R. Yager, and L. Zadeh, editors, *Fuzzy Logic and Soft Computing*. World Scientific Press, 1995.

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

[16] R. R. Yager. Toward a language for specifying summarizing statistics. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 33(2):177–187, 2003.

[17] R. R. Yager. A human directed approach for data summarization. In *IEEE International Conference on Fuzzy Systems*, pages 707–712, 2006.