

Linguistic Local Change Comparison Of Time Series

Rita Castillo-Ortega and Nicolás Marín
University of Granada
Granada, Spain
Email: {rita,nicm}@decsai.ugr.es

Daniel Sánchez
European Centre for Soft Computing
Mieres, Asturias, Spain
Email: daniel.sanchezf@softcomputing.es

Abstract—In this paper a methodology is proposed for providing a linguistic summary of the comparison of time series. This series are obtained as a result of queries which the user defines over a data cube with time dimension. Our objective is that the final summary describe the similarity between series by using local changes. Definitions of the degree and sign of local changes are proposed. The understandability of the linguistic summary is achieved by means of linguistically quantified sentences. So are the coverage and brevity by using a hierarchical partition of time dimension inherited from the data warehouse. Our linguistic summaries are well suited to be included in an interface layer of a data warehouse system, improving the quality of human-machine interaction and the understandability of the results.

Index Terms—Linguistic Summarization, Time series, OLAP, Business Intelligence, Multidimensional Data Model, Fuzzy Logic.

I. INTRODUCTION

Summarization is widely accepted as a data mining task, since summaries of data are extracted non-trivially, they are very useful since they contribute to a better understanding, and they provide a novel, previously unseen view of data.

In the setting of data mining, summaries are expected to be an understandable description of knowledge present in the data. The simpler form of such knowledge are plain facts, as in the case of using statistics (i.e., the average value of a certain variable in a database, etc.). However, much complex summaries may be obtained using more sophisticated knowledge representation languages with richer semantics. Due to its good understandability to the user, one very interesting possibility is to obtain linguistic summaries of data, i.e., summaries expressed in (usually a subset of) natural language.

One of the areas where linguistic summaries can be very useful is that of Business Intelligence solutions, intended to help managers in order to obtain a better understanding of its commercial procedure and operations and, this way, supporting decision-making. Data warehousing systems use to incorporate an interface layer responsible of providing such information (see for example the case of scorecards [1]). One of the tools that undoubtedly can help in building this kind of interfaces is to use natural language as a vehicle to describe data, as a human assistant in charge of this task would do [2].

Part of the research was supported by the Andalusian Government (Junta de Andalucía, Consejería de Innovación, Ciencia y Empresa) under project P07-TIC-03175 *Representación y Manipulación de Objetos Imperfectos en Problemas de Integración de Datos: Una Aplicación a los Almacenes de Objetos de Aprendizaje*. Part of the research was supported by the Spanish Government (Science and Innovation Department) under project TIN2009-08296.

Many queries to data warehousing systems involve the evolution of data along time. Data cubes usually have a time dimension because time plays a crucial role in any commercial activity. Hence, many OLAP queries in data cubes with time dimension produce as result time series data. In particular, many queries are intended to provide a comparison between datasets, like for example the total amount of sales in different shops of a supermarket chain or, involving time, about the evolution of those sales in the same shops. As a consequence, there is a necessity for tools able to provide understandable, accurate, and relevant information about the similarities and differences between time series.

In this work we propose a methodology for providing a linguistic summary of the comparison of time series obtained as a result to a user's query. The summary comes in the form of a collection of *accurate* linguistic quantified sentences describing the similarity between local changes in the whole time domain considered. On the basis of a hierarchical partition of the time dimension, available through the datacube structures, we not only achieve a complete *coverage* of the time domain, but we also avoid obtaining two sentences for the same time period and, in this sense, our summary is also *brief*. The *relevance* of the information is assumed since it is the user who asks for the comparison of the series by means of a specific query, and since we choose the most representative among all the possible sentences for a given time period, also on the basis of the user's preferences. Finally, as we indicated before, *understandability* is taken into account in that the information is provided in the form of a linguistic description in natural language.

As related work, many approaches to data summarization with the use of Soft Computing concepts can be found in the literature. Among these proposals, linguistic summarization is of special interest because of the above mentioned need for sentences close to natural language to describe data. An important seminal part of this research can be found in Ronald R. Yager's works, where he first proposes the use of quantified sentences in the sense of L. Zadeh [3] and later introduces his own way of quantification with OWA (Ordered Weighted Averaging) operators in [4], [5]. Also following Zadeh's footsteps, we can find J. Kacprzyk et al. ([6], [7], [8], [9]) introducing new quality measures and using the protoform concept in order to reach different summary profiles. The same author

The query language employed is not relevant for the purpose of our work.

has provided methods for linguistic comparison of time series [10], though of a different kind and with different objectives than we intend here. Raschia et al. have created a model named SaintEtiQ [11], working with hierarchies composed by summaries with different abstraction levels. From a different point of view, Bosc and Dubois propose the use of association rules to obtain knowledge from data and even for case based prediction [12]. Our work here is based on an approach to the linguistic summarization of a single time series on the basis of a hierarchical partition of time dimension [13]. The basis for a better understanding of this work can be found in [14], [15] and a detailed study is available at [16].

The document is organized as follows: Section II is devoted to present the source our data come from, introducing OLAP queries on fuzzy data cubes as a mean to obtain the time series to compare. The use of fuzzy data cubes with time dimension provides us with a hierarchical organization of the time domain that plays a determining role in our approach. Section III introduces our proposal to compare two query results in the form of two time series that share the same time line. In this section, we first analyze the characteristics of the series where the comparison focus on and then we set the linguistic framework used to build the summaries. Two greedy algorithms to compute the linguistic summaries are also presented in this section. Section IV contains a descriptive example of the use of the proposed approach and, finally, section V concludes the paper.

II. DESCRIBING TIME SERIES COMING FROM OLAP CUBES

Time plays a main role in data warehouses that store *historical* information regarding the daily operation of the organization. This way, decision-makers can analyze the evolution of several aspects of the business along the time. One of the most extended models for building data warehouses is the use of datacubes (also known as OLAP cubes).

We can briefly describe a data cube as a n -dimensional set of cells. Each dimension consists of a set of members that can be arranged along a hierarchy or many hierarchies. Thanks to the hierarchical organization of dimensions, granularity in the data cube can vary according to the selected level chosen in each dimension. OLAP operations permit the user to take advantage of this fact and query the data cube in order to get the desired data in each moment. According to the previous paragraph, among the dimensions of a data cube, we usually find the time dimension.

OLAP cubes have been extended in order to add the capability of dealing with fuzzy data. In our approach, we use as basis the model proposed in [17], [18]. One of the main characteristics of this model is that the hierarchical organization of dimensions can be made by means of linguistic partitions. The use of this model provides a suitable framework for the linguistic description of time series.

Figure 1 describes a (generic) data cube with a time dimension. Each cell contains the value of the variable V according to the context defined by the members of each dimension. As we can see, using conventional OLAP operations, we can query

the data cube focusing on a given member of each of the non-time dimensions: the result of this query is a time series of data.

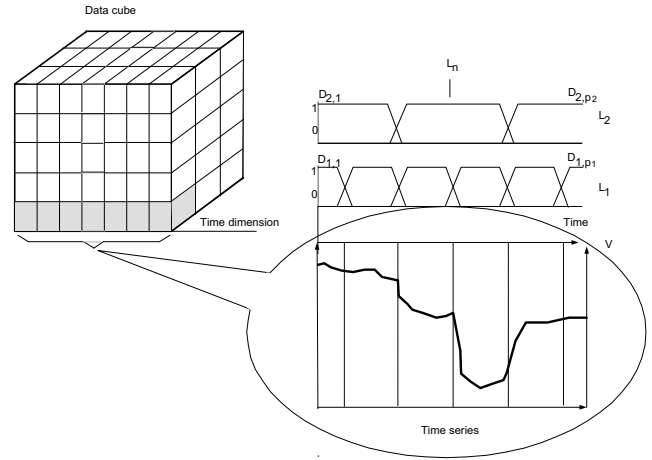


Fig. 1. Time series from OLAP cubes

Suppose that the time dimension is described in its finest grained level of granularity by members $T = \{t_1, \dots, t_m\}$. Then, the above mentioned query result will have the following form: $\{ \langle t_1, v_1 \rangle, \dots, \langle t_m, v_m \rangle \}$, where every v_i is a value of the basic domain of the variable V . We will use $TS(t_i)$ to refer to v_i when necessary.

In order to linguistically describe the time information in our approach, the time dimension is hierarchically organized in n levels, namely, $L = L_1, \dots, L_n$. Each level L_i has a partition $\{D_{i,1}, \dots, D_{i,p_i}\}$ of the basic time domain associated. 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 $\{X_1, \dots, X_r\}$ is a partition on X iff:

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

Additionally, considering 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).

As an example, consider a data warehouse with data related to medical centers in a given country. A data cube of this data warehouse can be built with information regarding the patient inflow according to the dimensions *medical center*, *service*, and *time*. If we focus (slice and dice) on a given medical center C and aggregate (roll up) across the service dimension, we obtain the time series “Total patient inflow in center C

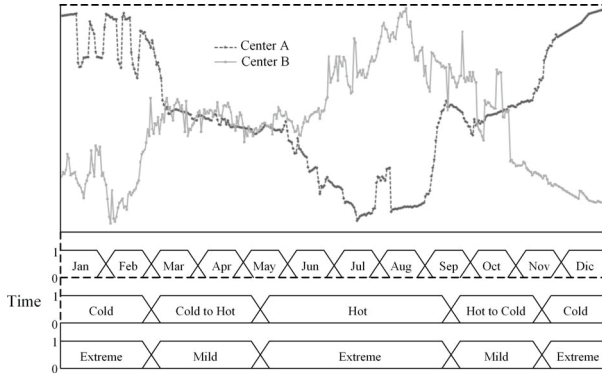


Fig. 2. Patient inflow data series

along the time”. In this example, the time dimension is fuzzy hierarchically organized using a meteorological criteria.

Figure 2 contains two time series that shows the patient inflow along a given year to two medical centers. As we can see, the time dimension is hierarchically organized thanks to three fuzzy partitions 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 based on a meteorological criteria with two levels of granularity. Fuzziness is specially useful in these two last partitions because transitions between periods are clearly fuzzy.

In previous works, we have focused on the linguistic summarization of a single time series [13]. In this work, we extend our approach to the case of the linguistic comparison of the trends of two time series like the ones depicted in Figure 2.

III. LOCAL CHANGE BASED COMPARISON OF TIME SERIES

Once we have described the framework where the time series are defined, the next step is to introduce the new approach to linguistically compare the information related to two time series. There are several ways to compare series. Among them, special attention has been paid to value-based comparison and trend-based comparison. In a previous work [13] we introduced an approach to value-based comparison of two time series. With respect to trend-based comparison, Kacprzyk et al.’s work is of interest [10]. This approach compares the trends through the analysis of significant segments in the series with respect to duration, dynamics (slope), and variability.

In this work, our interest is centered on the comparison of trends but with a different perspective: we study the dynamics of change through the local variations in the points of the series and we take advantage of the hierarchical partition of the time domain when building the summaries. Our summaries are comprised of several linguistically quantified sentences.

A. Comparing the dynamics of change

The current subsection is devoted to describe the process we follow in order to compare the dynamics of change. The bases

will be set by means of the introduction of some important definitions.

We use consecutive points in order to analyze the dynamics of change in the series. To do that, we measure the local change for every time point with respect to the next one in each individual series. This local change can be formalized with the next definitions.

Definition 1 (Local change): Let $TS = \{ \langle t_1, v_1 \rangle, \dots, \langle t_m, v_m \rangle \}$ be a time series and t_i with $i \in [1, m-1]$ be a given time point. The local change $\Delta_{TS}(t_i)$ of t_i in TS is defined as follows: $\Delta_{TS}(t_i) =$

$$= \arctan \left(\frac{TS(t_{i+1}) - TS(t_i)}{t_{i+1} - t_i} \right) = \arctan \left(\frac{v_{i+1} - v_i}{t_{i+1} - t_i} \right)$$

Hence, we are defining local change as the slope of the line joining two consecutive points in the time series. We assume that this slope comes in the form of a value in $[-\pi/2, \pi/2]$.

In Figure 3 the reader can graphically see the intuitive semantics behind our definitions concerning the local change at a given point t_i .

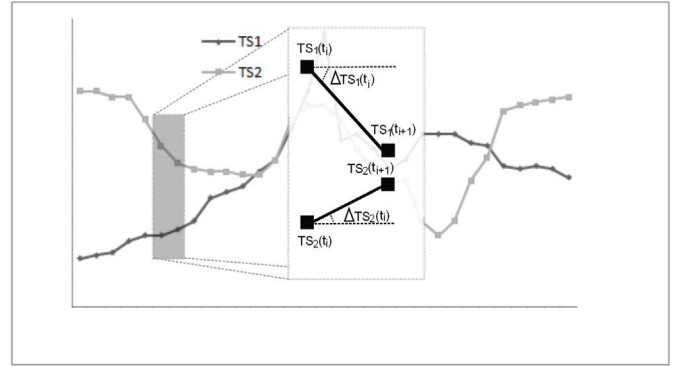


Fig. 3. Local change

Our approach tries to linguistically compare the local change of two time series. To do that, we need a linguistic framework in order to express this comparison by means of natural language. The following definitions are devoted to this aim.

Definition 2 (Same sign function): Let TS_1 and TS_2 be two time series defined on $t_1..t_m$. The same sign function SS_{TS_1, TS_2} is defined, for every $t_i \in \{t_i, \dots, t_m-1\}$ as follows:

$$SS_{TS_1, TS_2}(t_i) = \begin{cases} 1, & \text{if } (\Delta_{TS_1}(t_i) * \Delta_{TS_2}(t_i) \geq 0) \\ 0, & \text{otherwise} \end{cases}$$

Similarly, we can define the different sign function.

Definition 3 (Different sign function): Let TS_1 and TS_2 be two time series defined on $t_1..t_m$. The different sign function DS_{TS_1, TS_2} is defined, for every $t_i \in \{t_i, \dots, t_m-1\}$ as follows:

$$DS_{TS_1, TS_2}(t_i) = \begin{cases} 0, & \text{if } (\Delta_{TS_1}(t_i) * \Delta_{TS_2}(t_i) \geq 0) \\ 1, & \text{otherwise} \end{cases}$$

It is obvious that $SS_{TS_1, TS_2}(t_i) = 1 - DS_{TS_1, TS_2}(t_i)$. If we only focus on the magnitude of the change, we can define two additional functions.

Definition 4 (Degree of similar variation): Let TS_1 and TS_2 be two time series defined on $t_1..t_m$ and t_i with $i \in [1, m-1]$ be a given time point. The degree of similar variation $SV_{TS_1, TS_2}(t_i)$ in point t_i between series TS_1 and TS_2 is defined as follows:

$$SV_{TS_1, TS_2}(t_i) = 1 - \frac{||\Delta_{TS_1}(t_i)| - |\Delta_{TS_2}(t_i)||}{\pi/2}$$

By means of the computation of the degree of similar variation we obtain the similarity observed between the two angles responsible of the local change at a given t_i in both series, without taking into account the sign.

Similarly, we can define the degree of different variation:

Definition 5 (Degree of different variation): Let TS_1 and TS_2 be two time series defined on $t_1..t_m$ and t_i with $i \in [1, m-1]$ be a given time point. The degree of different variation $DV_{TS_1, TS_2}(t_i)$ in point t_i between series TS_1 and TS_2 is defined as follows:

$$DV_{TS_1, TS_2}(t_i) = \frac{||\Delta_{TS_1}(t_i)| - |\Delta_{TS_2}(t_i)||}{\pi/2}$$

Again, it is obvious that $SV_{TS_1, TS_2}(t_i) = 1 - DV_{TS_1, TS_2}(t_i)$.

B. Linguistically quantified sentences

Fuzzy quantification extends classical quantification by considering (fuzzy) linguistic quantifiers, a generalization of the ordinary quantifiers $\exists$ and $\forall$ of first order logic [19]. A large number of applications can be found in the literature in areas like quantifier-guided aggregation, linguistic summarization, computing with words, and quantification in fuzzy description logics, among many others.

The most usual quantified sentences considered in the literature are of the form “Q of X are A” or “Q of D are A”, where Q is a linguistic quantifier, X is a (finite) crisp set, and A, D are fuzzy subsets of X. These sentences are called type I and type II sentences, respectively. Linguistic quantifiers are normal, convex fuzzy subsets of integers $\mathbb{Z}$ (absolute quantifiers) or $[0, 1]$ (relative quantifiers). Particularly, in this paper, linguistic quantifiers are represented by trapezoidal functions.

There are many different approaches for evaluating quantified sentences; we shall use the method called *GD* introduced in [20] as follows: the evaluation of “Q of D are A” by means of GD is

$$GD_Q(A/D) = \sum_{\alpha_i \in \Delta(A/D)} (\alpha_i - \alpha_{i+1}) Q\left(\frac{|(A \cap D)_{\alpha_i}|}{|D_{\alpha_i}|}\right) \quad (1)$$

where $(A \cap D)(x) = \min(A(x), D(x))$, $\Delta(A/D) = \Delta(A \cap D) \cup \Delta(D)$, $\Delta(D)$ being the level set of D , and $\Delta(A/D) = \{\alpha_1, \dots, \alpha_p\}$ with $\alpha_i > \alpha_{i+1}$ for every $i \in \{1, \dots, p\}$, we consider $\alpha_1 = 1$ and $\alpha_{p+1} = 0$ (although α_{p+1} is not in

the level set we consider it in the formula). The set D is assumed to be normalized. If not, D is normalized and the same normalization factor is applied to $A \cap D$ (D and $A \cap D$ will be divided by the greater value in D).

The GD method has been used due to its efficiency and non-strict character. The method also fulfills some interesting properties related to relative quantifiers (defined in [20]). Another point is that it is easy to implement. Anyway, let us remark that the strategies presented in this paper are not dependent on the evaluation method.

As we have mentioned, we consider that the user is interested in linguistic comparisons that take the form of a collection of quantified sentences. In this sense, our approach will deliver a collection of sentences of the form “Q of $D_{i,j}^S$ are A^S ” where:

- Q is a linguistic quantifier.
- $D_{i,j}$ is a label member of a certain level i of the hierarchy associated to the time dimension and

$$D_{i,j}^S(< t, v >) = D_{i,j}(t). \quad (2)$$

- A is a label or the union of a subset of labels of the partition of the variable V under study. In this case V represents the local change, and

$$A^S(< t, v >) = A(v). \quad (3)$$

With this kind of sentences, the approach will be able to produce results like “Most (Q) days of the cold season ($D_{i,j}^S$), both series exhibit a local change with the same sign and similar variation (A^S)”.

With respect to the quantifier Q , 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 $\{Q_1, \dots, Q_{qmax}\}$ of a coherent family of quantifiers $\mathcal{Q}$ [21] to be used in the summarization process.

Definition 6 (Coherent family of quantifiers): Let $\mathcal{Q} = \{Q_1, \dots, Q_l\}$ be a linguistic quantifier set, we shall say it is coherent if it verifies that:

- The membership functions of elements in $\mathcal{Q}$ are non-decreasing functions.
- A partial order relation $\succeq$ is defined in $\mathcal{Q}$. It has as its maximal element $Q_1 = \exists$ and as its minimal one $Q_l = \forall$. Furthermore $\forall Q_i, Q_j \in \mathcal{Q}, Q_i \subseteq Q_j \Rightarrow Q_j \succeq Q_i$.
- The membership function of the quantifier $\exists$ is given by $Q_1(x) = 1$ if $x \neq 0$ and $Q_1(0) = 0$, whereas the membership functions of $\forall$ will be $Q_l(x) = 0$ if $x \neq 1$ and $Q_l(1) = 1$.

To play the role of D , we can take any of the labels $D_{i,j}$ defined in the hierarchy associated to the time domain in our fuzzy OLAP cube.

Finally, we can consider the definition of A according to the previous definitions:

As the data set X is assume to be finite, D is considered to be finite, and hence the number of relevant α -cuts is also finite.

- Points where local change has the same sign (whatever the variation). In this case:

$$A(t_i) = SS_{TS_1, TS_2}(t_i)$$

- Points where local change has different sign (whatever the variation). In this case:

$$A(t_i) = DS_{TS_1, TS_2}(t_i)$$

- Points where local change has similar variation (whatever the sign). In this case:

$$A(t_i) = SV_{TS_1, TS_2}(t_i)$$

- Points where local change has different variation (whatever the sign). In this case:

$$A(t_i) = DV_{TS_1, TS_2}(t_i)$$

C. Obtention of the summaries

According the above defined linguistic framework, our approach will deliver a summary composed by a collection of sentences of the form “ Q of D^S are A^S ” with the above described semantics. To obtain this collection, apart from the linguistic framework analyzed in the previous subsection, the user will provide a threshold τ for the minimum accomplishment degree she/he wishes for the quantified sentences comprising the summaries. It is also necessary that the user specify whether she/he is interested in summarizing the local changes according to the sign or the variation.

Once we have defined all the parameters we need, it is time to launch a process able to compute the desired summary. Algorithm 1 explores all the possible A labels of C before exploring solutions with a less strict quantifier, beginning with the *strictest* quantifier provided by the user.

In order to look for *brevity*, we start from the time periods at the top level of the hierarchy. Each level has its own quantifier bound ($Qbound_i$) that indicates the least strict quantifier to be considered. The set *ToSummarize* is the collection of time periods for which a quantified sentence is missing. The set *Summarized* is the collection of time periods for which a quantified sentence is found. If it is possible to obtain an accomplishment degree greater than or equal to τ for a certain period using a quantifier Q and an A label of C , the procedure obtains a summary for that period.

If all A labels are tried without success, the algorithm repeats the process again, but with a less strict quantifier, until $Qbound_i$ is reached. When a summary is found in a certain time period we say that the period is *covered*.

If there is no result 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 of the time domain (line 21).

For every $D_{i,j}$, $ch(D_{i,j})$ is defined as follows:

- $ch(D_{1,j}) = \emptyset$ for all j .
- Otherwise, $ch(D_{i,j}) = \{D_{i-1,k}, k \in \{1..p_{i-1}\} | D_{i-1,k} \cap D_{i,j} \neq \emptyset \text{ and } \neg \exists D \in ToSummarize \cup Summarized, (D_{i-1,k} \cap D_{i,j}) \subseteq D\}$.

Algorithm 1 to obtain linguistic summaries.

Input

Two time series TS_1 and TS_2
 A hierarchical fuzzy partition of the time dimension D
 A totally ordered subset $\{Q_1..Q_{qmax}\}$ of a coherent family of quantifiers Q .
 A threshold τ as minimum accomplishment for quantified sentences.
 For each L_i , maximum number of quantifiers to use ($Qbound_i$).
 Two collections of alternatives for A , $C \subseteq \{SS_{TS_1, TS_2}, DS_{TS_1, TS_2}\}$ or $C \subseteq \{SV_{TS_1, TS_2}, DV_{TS_1, TS_2}\}$.

Output

A Summary of the comparison of time series TS_1 and TS_2

Algorithm

```

1: ToSummarize  $\leftarrow L_n$ ;
2: Summary  $\leftarrow \emptyset$ ; Summarized  $\leftarrow \emptyset$ ;
3: while ToSummarize  $\neq \emptyset$  do
4:   Take  $D_{i,j} \in ToSummarize$ 
5:   ToSummarize  $\leftarrow ToSummarize \setminus \{D_{i,j}\}$ ;
6:    $p \leftarrow qmax$ ; covered  $\leftarrow false$ ;
7:   while  $p \geq Qbound_i$  and not covered do
8:     Let  $A \leftarrow argmax_{B \in C_k} GD_{Q_p}(B^S/D_{i,j}^S)$ ;
9:     if  $GD_{Q_p}(B^S/D_{i,j}^S) \geq \tau$  then
10:      Summary  $\leftarrow Summary \cup \{Q_p \text{ of } D_{i,j}^S \text{ are } A^S\}$ ;
11:      Summarized  $\leftarrow Summarized \cup (D_{i,j})$ ;
12:      covered  $\leftarrow true$ ;
13:   end if
14:    $p \leftarrow p - 1$ ;
15: end while
16: if not covered and  $i > 1$  then
17:   ToSummarize  $\leftarrow ToSummarize \cup ch(D_{i,j})$ ;
18: end if
19: end while

```

With regards to the computational complexity of the algorithms, as can be seen, it exponentially depends on $|A|$ (the number of labels defined in A) and linearly depends on $\sum_i p_i$ (the number of labels in the time partitions) and $qmax$ (the number of quantifiers). In any case, $|E|$ is usually low and it is not necessary to explore its entire powerset due to the user-defined bounds.

IV. EXPERIMENTATION

This section is devoted to show the performance of the proposed approach. With that purpose we use two series representing the gasoline price in New York and Texas during the same period of time.

As it was mentioned before, once the user has the corresponding time series data, he/she is in charge of defining the linguistic background and setting the parameters following his/her specific needs. The different partitions on the variable and time domains will define the type of sentences he/she will be able to obtain at the end of the summarization process. Moreover, he/she has to set the parameters values to inform the algorithm of his/her exploration preferences (sign/variation, threshold and boundaries).

The series used in this example represent the weekly all grades all formulation retail gasoline prices (in cents per gallon) in New York and Texas during the decade 2000-2010. In this example we do not have all the measures on year 2000 but only those corresponding to the second half of the year. The same thing occurs when dealing with 2010 where we only have

values for the first quarter of the year. In total, we count on 513 measures representing the 513 weeks. Series are depicted in Figure 4.

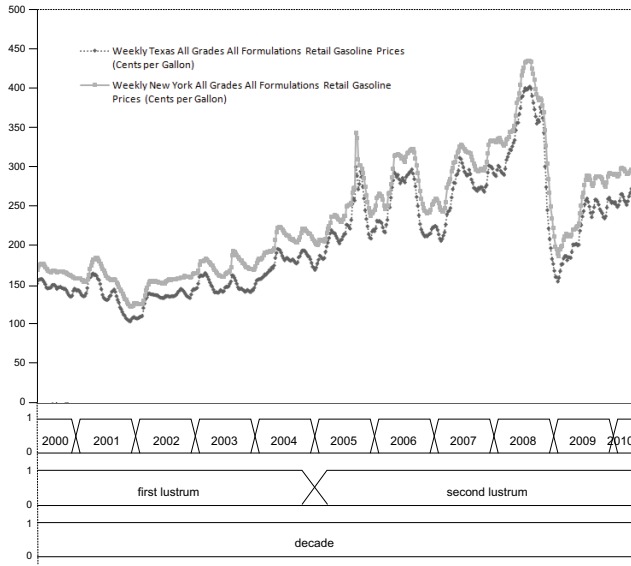


Fig. 4. Weekly all grades all formulation retail gasoline prices (in cents per gallon)

As we can see in Figure 4, the time dimension is hierarchically organized in three levels. The more general level represents the whole decade, while the second is made up of partitions representing two approximate periods of five years (lustrums). Finally, the finest grained level partition represents approximate years. In this context, fuzziness is used in order to avoid sharp transitions between time periods. The hierarchical partition of the time domain is defined in Table I.

Level	Label	Definition
1	Second half of year 2000	(0, 0, 28, 32)
	Year 2001	(28, 34, 80, 84)
	Year 2002	(80, 84, 132, 136)
	Year 2003	(132, 136, 184, 188)
	Year 2004	(184, 188, 236, 240)
	Year 2005	(236, 240, 288, 292)
	Year 2006	(288, 292, 341, 345)
	Year 2007	(341, 345, 393, 397)
	Year 2008	(393, 397, 445, 449)
	Year 2009	(445, 449, 497, 501)
	First quarter of year 2010	(497, 501, 513, 513)
2	First lustrum	(0, 0, 230, 249)
	Second lustrum	(230, 249, 513, 513)
3	Whole decade	(0, 0, 513, 513)

TABLE I
PARTITION OF THE TIME DIMENSION

The example has been carried out using the set of linguistic quantifiers shown at Table II. The threshold value has been set as $\tau = 0.7$. Regarding $Qbound_i$, this parameter has value 1 in levels 3 and 2, and value 2 in level 1 ($Qbound_3 = Qbound_2 = 1$ and $Qbound_1 = 2$). This is due to the fact that the user wants

to use stricter quantifiers in levels holding bigger periods of time, and less strict quantifiers in the finest grained partition in order to obtain representative summaries.

Quantifier	Definition
Most	(0, 0.7, 0.9, 1)
At least 70%	(0, 0.6, 0.8, 1)
At least 60%	(0, 0.5, 0.7, 1)
At least 50%	(0, 0.4, 0.6, 1)

TABLE II
QUANTIFIERS

In the first case, we are going to consider that the user is interested or she/he considers useful only those labels representing the local changes in terms of sign as C . As a result, we obtain this set of quantified sentences.

- Most days of second lustrum, both series exhibit a local change with the same sign (0.716765).
- At least 70% of the days of a half of year 2000, both series exhibit a local change with the same sign (0.770035).
- Most days of year 2001, both series exhibit a local change with the same sign (0.714104).
- At least 70% of the days of year 2003, both series exhibit a local change with the same sign (0.821261).
- Most days of year 2004, both series exhibit a local change with the same sign (0.812402).

In the second case, the user focusses the attention on those labels representing the local changes in terms of variation as C . The obtained set of quantified sentences is:

- At least 70% of the days of a half of year 2000, both series exhibit a local change with similar variation (0.727618).
- At least 70% of the days of year 2002, both series exhibit a local change with similar variation (0.707482).
- At least 70% of the days of year 2003, both series exhibit a local change with similar variation (0.763968).
- At least 70% of the days of year 2004, both series exhibit a local change with similar variation (0.784934).
- Most days of year 2005, both series exhibit a local change with similar variation (0.761454).
- Most days of year 2006, both series exhibit a local change with similar variation (0.736421).
- At least 70% of the days of year 2007, both series exhibit a local change with similar variation (0.75718).
- Most days of year 2008, both series exhibit a local change with similar variation (0.736357).
- At least 70% of the days of year 2009, both series exhibit a local change with similar variation (0.754698).
- It has not been possible to find any kind of information referring to the rest of time periods.

The reader can see that the algorithm has obtained a short summary using sign local changes while the summary is a little bit longer using variation local changes. In the first execution the algorithm has come up with a suitable sentence describing the whole second lustrum. In the second execution

the accuracy of the sentence produce for the same period of time using variation is not enough and smaller periods have to be analyzed. We can conclude that the series are more similar according to local changes in sign that according to local changes related to variation.

We have to keep in mind that the presented method analyzes local trends by doing the calculation with pairs of consecutive time points. It could be possible that the obtained results surprised the user if he/she was expecting a summary comprising a global trend description. In some occasions the local trends can coincide with the global one, but this is not necessarily true.

V. CONCLUSION AND FUTURE WORK

Linguistic summarization of data is a very interesting data mining task providing highly understandable knowledge. In this work we have proposed a methodology for providing linguistic summaries to answer a usual kind of queries in data warehouses, i.e., the comparison of time series. These queries are frequent as data cubes use to have a time dimension.

The comparison is focused on the analysis of the degree and sign of the local change observed in every time point. The summary is a collection of *accurate* linguistic quantified sentences describing the similarity between local changes in the whole time domain considered. A hierarchical partition of the time dimension permits us not only to achieve a complete *coverage* of the time domain, but also to obtain *brief* summaries. The *relevance* of the information is assumed since it is the user who asks for the comparison of the series by means of a specific query, and since we choose the most representative among all the possible sentences for a given time period, also on the basis of the user's preferences. Our linguistic summaries are well suited to be included in an interface layer of a data warehouse system, improving the quality of human-machine interaction and the understandability of the results.

We are currently extending the approach so that the user can interleave both the analysis in term of sign and in term of variation at the same time. This way, different levels of granularity can be managed by the algorithm in its look for a *brief* and *accurate* summary of the comparison of the dynamic of local changes in the series.

REFERENCES

- [1] R. S. Kaplan and D. P. Norton, *The Balanced Scorecard: Translating Strategy into Action*. Harvard Business Press, 1 ed., September 1996.
- [2] F. Bobillo, M. Delgado, J. Gómez-Romero, and E. López, "A semantic fuzzy expert system for a fuzzy balanced scorecard," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 423–433, 2009.
- [3] R. R. Yager, "A new approach to the summarization of data," *Information Sciences*, no. 28, pp. 69–86, 1982.
- [4] R. R. Yager, "Toward a language for specifying summarizing statistics," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 33, no. 2, pp. 177–187, 2003.
- [5] R. R. Yager, "A human directed approach for data summarization," in *IEEE International Conference on Fuzzy Systems*, pp. 707–712, 2006.
- [6] J. Kacprzyk, "Fuzzy logic for linguistic summarization of databases," in *IEEE International Fuzzy Systems Conference*, pp. 813–818, 1999.
- [7] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," in *International Journal of General Systems*, vol. 30, pp. 133–154, 2001.
- [8] 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*, vol. 10, pp. 813–834, 2000.
- [9] J. Kacprzyk and S. Zadrozny, "Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools," *Inf. Sci. Inf. Comput. Sci.*, vol. 173, no. 4, pp. 281–304, 2005.
- [10] 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 *IFSA-EUSFLAT 2009* (U. K. J. P. Carvalho, D. Dubois and J. M. C. Sousa, eds.), pp. 1321–1326, 2009.
- [11] G. Raschia and N. Mouaddib, "Saintetiq: a fuzzy set-based approach to database summarization," *Fuzzy Sets Syst.*, vol. 129, no. 2, pp. 137–162, 2002.
- [12] P. Bosc, D. Dubois, O. Pivert, H. Prade, and M. D. Calmes, "Fuzzy summarization of data using fuzzy cardinalities," in *Int. Conf. Inf. Process. Manag. Uncertainty Knowl. Based Syst.*, pp. 1553–1559, 2002.
- [13] R. Castillo-Ortega, N. Marín, and D. Sánchez, "Time series comparison using linguistic fuzzy techniques," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2010*, pp. 330–339, 2010.
- [14] R. Castillo-Ortega, N. Marín, and D. Sánchez, "Fuzzy quantification-based linguistic summaries in data cubes with hierarchical fuzzy partition of time dimension," in *IDEAL'09* (H. Yin and E. Corchado, eds.), vol. 5788 of *LNCS*, pp. 578–585, Springer, Heidelberg, 2009.
- [15] R. Castillo-Ortega, N. Marín, and D. Sánchez, "Linguistic summary-based query answering on data cubes with time dimension," in *FQAS'09* (T. Andreasen, R. R. Yager, H. Bulskov, H. Christiansen, and H. L. Larsen, eds.), vol. 5822 of *LNAI*, pp. 560–571, Springer, Heidelberg, 2009.
- [16] R. Castillo-Ortega, N. Marín, and D. Sánchez, "A fuzzy approach to the linguistic summarization of time series," *Journal of Multiple-Valued Logic and Soft Computing, Special Issue Soft Computing Techniques in Data Mining*, vol. 17, no. 2,3, pp. 157–182, 2011.
- [17] N. Marín, C. Molina, J.-M. Serrano, and M. A. V. Miranda, "A complexity guided algorithm for association rule extraction on fuzzy data cubes," *IEEE T. Fuzzy Systems*, vol. 16, no. 3, pp. 693–714, 2008.
- [18] C. Molina, L. R. Ariza, D. Sánchez, and M. A. V. Miranda, "A new fuzzy multidimensional model," *IEEE T. Fuzzy Systems*, vol. 14, no. 6, pp. 897–912, 2006.
- [19] L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages," *Computing and Mathematics with Applications*, vol. 9, no. 1, pp. 149–184, 1983.
- [20] M. Delgado, D. Sánchez, and M. Vila, "Fuzzy cardinality based evaluation of quantified sentences," *International Journal of Approximate Reasoning*, vol. 23, pp. 23–66, 2000.
- [21] 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 *Fuzzy Logic and Soft Computing* (B. Bouchon-Meunier, R. Yager, and L. Zadeh, eds.), World Scientific Press, 1995.