16th World Congress of the International Fuzzy Systems Association (IFSA)
9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)

Linguistic Summarization of Processes – a research agenda

Anna Wilbik Uzay Kaymak

Information Systems, School of Industrial Engineering,
Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: A.M.Wilbik@tue.nl and U.Kaymak@tue.nl

Abstract

Linguistic summarization offers novel ways of gaining insight into large amounts of data by extracting their main properties and representing them linguistically. Various linguistic summarization techniques have been proposed for data sets consisting of attribute-value pairs. However, many application domains are characterized by structured and relational data, where explicit relations amongst data elements exist. Process data, in which business activities are ordered sequentially, is one example. Linguistic summarization for such data has been considered only sparsely in the literature. In this paper, we consider the challenges for obtaining linguistic summaries from process data and propose a research agenda.

Keywords: Linguistic summaries, process mining, event logs, linguistic process models.

1. Introduction

As larger and larger amounts of data are collected and stored through contemporary information systems, making sense of these large amounts of data and generating understanding becomes more and more important. Data visualization is often used to increase the understanding of the data by emphasizing or discovering visual cues that help humans discover relations in the data that would otherwise have remained invisible or difficult to notice [23].

Another way of generating insight into the data is summarization, where the most characteristic aspects of the data are presented linguistically in natural language [21]. This is also known as verbalization, which is another way of gaining insight into large amounts of data. It is complementary to visualization in that natural language is used to describe the main characteristics of the data instead of the graphical (visual) techniques of visualization. Because of the inherent imprecision of natural language, fuzzy set approaches to summarization have been studied by the fuzzy sets research community in order to deal with this imprecision.

In the literature, various linguistic summarization methods have been proposed based on the fuzzy set theory [11, 12, 1, 41, 30, 28]. Usually, these methods consider either time series data or data represented as attribute-value pairs. The early focus on these types of data is understandable, since they are very common and represent a large proportion of the data collected in various applications. Also, fields such as data mining and machine learning initially analyzed these types of data. Nevertheless, there are many application domains where the data has a natural structure consisting of (pre-defined) relations amongst data elements. In this paper, we use the term *structured data* for denoting such relational data. Data containing meta-information making explicit the relations amongst data elements, semantically enriched data, etc. fall under this category. A common representation for structured data is a graph, where the nodes and the edges in the graph correspond to the structure of the data. We contend that new linguistic summarization methods are needed for dealing with the structured data.

An important domain in which structured data is encountered often is the domain of (business) process modeling. For organizations and businesses, process modeling is the basis for business process management that helps an organization to better control its business processes and hence increase customer value and operational excellence. Process data typically consists of a series of activities executed in parallel or sequentially, depending on the specific process semantics and context. Today, information systems can make a log of these activities in the so-called event logs, which can be used to determine the main characteristics of a business' processes. Oftentimes, the processes are very complex and show a large degree of variation across different cases. However, process models should reflect the main aspects of a business process and generalize from the ad hoc deviations that may occur. Linguistic summarization is potentially a good way of extracting this description from process data.

However, summarization of process data has not received much attention in the literature. In this paper, we consider linguistic summarization of process data based on fuzzy set-based methods. The problem poses many challenges. We identify the main ones and illustrate some of them with examples. Based on this exposition, we also propose a research agenda to bring the field forward.

The outline of the paper is as follows. In Section 2 we present a brief overview of previous work in lin-

guistic summarization. We discuss process discovery from logged data in Section 3. In Section 4, we present the challenges that we have identified for the fuzzy set community for processing event logs and generating linguistic summaries of processes. This leads to an overall research agenda that we present in Section 5. Finally, the conclusions are given in Section 6.

2. Linguistic summarization

The idea of data summarization was strongly investigated by many researchers, *e.g.,* Bosc [4], Dubois and Prade [11, 12], Keller and his collaborators [1, 2, 3, 31, 38], Yager, Kacprzyk and their collaborators [17, 19, 30, 41, 42], Wu and Mendel [40], Marin and Sanchez [7, 8], Raschia and Mouaddib [28, 29], Bouchon-Meunier and her collaborators [6, 24, 25], Mendel [27]. There are many different approaches, but it seems the most commonly used is the one by Yager [41], that was implemented in many different applications, *e.g.,* retail (*e.g.,* Kacprzyk *et al.* [14]), finance (*e.g.,* Castillo-Ortega *et al.* [8]), eldercare (*e.g.,* Ros *et al.* [31] or Wilbik *et al.* [38, 39]). In this approach $Y = \{y_1, y_2, \ldots, y_n\}$ is the set of objects (records) in the database D, e.g., a set of employees, and $A = \{A_1, A_2, \ldots, A_m\}$ is the set of attributes (features) characterizing objects from Y, e.g., a salary, age in the set of employees.

A linguistic summary is a template-based quantified proposition built from the following elements:

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

There are two main protoforms (templates): simple protoform (Q y's are P), e.g., "most employees earn low salary" and extended form, including a qualifier (Q R y's are P), e.g., "most young employees earn low salary".

Linguistic summaries besides summarizing the databases (*e.g.,* Kacprzyk *et al.* [20, 19]), were adapted also for other types of data, such as time series (Kacprzyk *et al.* [16], Castillo-Ortega *et al.* [8]), texts (*e.g.,* Szczepaniak [32]), videos (*e.g.,* Anderson *et al.* [1, 2, 3]), sensor data (*e.g.,* Ros *et al.* [31], Wilbik *et al.* [38, 39]), web logs (*e.g.,* Zadrożny and Kacprzyk [45]) and recently event logs (Wilbik and Kaymak [37]). However so far none of the authors were focusing on the sequences of actions or events.

In this paper we focus on the last type of data mentioned here, namely event logs. They are traces of executing some (business) processes by the users. Such type of data is omnipresent, as most organization use information systems to support and execute their business processes [13] with work-flow management technology as a standard [34].

3. Generating Process Models

Fields such as industrial engineering aim to optimize complex processes of systems. Those processes can be recorded with event logs in which the observed event types are recorded together with the time stamp. In order to understand those processes and such data, methods were introduced or adapted in the domain of process mining. The goal of process mining is to extract non-trivial and useful process related information from event logs [33]. There are three traditional tasks of process mining:

- discovery – deriving information from some event log without using an a-priori model;
- conformance – checking if reality conforms to a model;
- extension – extending the a-priori model with a new aspect or perspective.

In this paper we focus only on the process discovery. Process mining analysis tries to find out the answer to several questions, such as which tasks or sequences of tasks are frequently executed, or what is the performance of the process in terms of throughput time, service time, waiting time or cost, or even, are there any specific properties related with worse or better performance. We believe that linguistic summarization may help to answer to such questions.

We think that in basic and general form the following analogy with the linguistic summaries of data can be assumed:

- $Y = \{y_1, y_2, \ldots, y_n\}$ is the set of cases (traces) in the event log, and
- $A = \{A_1, A_2, \ldots, A_m\}$ is the set of features characterizing those traces from Y, e.g., throughput time, presence of a sequence of actions, etc.

Then, similar protoforms as in the basic case may be obtained, allowing to construct summaries such as "most cases had a long throughput time" or in the extended form "most cases where action X took place, had a long throughput time".

Even though it looks pretty simple and same as in the original form, there are still some questions that need to be answered. Let us mention some of them: what are the features that characterize the traces and how to obtain them? Are Y only the cases (traces)? Are those summaries understandable by the domain experts? What is their usability?

In the next section we point out several challenges that are important to solve while adapting linguistic summaries for the process data. We discuss them and provide some examples.

4. Challenges

There are many challenges associated with summarizing the process data. We discuss here a few of those that are, in our opinion, most crucial.

4.1. Event logs

The first challenge is the data. The processes are usually recorded in the form of event logs. Event logs show occurrences of events at specific moments in time, where each event refers to a specific process and case [34]. Usually event logs are large text files with some standardized format, ordered according to the time.

Analysis of such data requires some preprocessing and data understanding. Even if we wish to compute the throughput time, we need first to group the events concerning the same case together, and then within those calculate the required values.

A single case within an observed process may have different number of actions. With certain actions different additional information may be included. As an example of an event log we may consider log of Volvo IT incident management system, that was used in BPI Challenge 2013 [35]. This log contains 7554 traces and 65533 events.

An excerpt of this log is shown in Table 1. Here we see 12 log entry for one case. Some fields in this case are repeated, as e.g., impact, product id or country owner. Finding in the file all actions related with one case or computing the throughput time requires some additional preprocessing.

However, the event logs may not always look so nice. Let us consider the health-care domain and a process of examining and healing the patient [22]. Very often in the information system there is no even proper event log, and the information about taken actions has to be very often extracted from the patient data. For instance if a certain test result, e.g., glucose level, was added to the patients medical history, then from this fact we can induce that this test was performed. However, such an activity may trigger other activities, not in a deterministic manner, but in a stochastic manner, since different patients react to the care activities in different ways (*e.g.,* a medicine can have different effects on different patients).

Also if we consider an event log where each action have start and termination time stamps, some actions may be performed parallel. This makes another challenge for the analysis.

There is a need for the suitable representation of such data, that will enable to store the events in a way that facilitates linguistic summarization analysis so that no laborious data preprocessing is necessary.

It would be also good to have an established framework or methodology on how to preprocess and prepare the event log data for further summarization analysis. It is also important to answer the question, which features are important to analyze. This brings us to the second challenge, that is related to the fact that a process is a sequence of actions.

4.2. Sequences

It is easy to create summaries such as, e.g., "in most cases action *Register* was present". However processes are sequences of actions. There were not many attempts to deal with sequences in linguistic summaries (see *e.g.,* Wilbik and Kacprzyk [36]). The analysis of sequences requires some further considerations.

One of the decisions that need to be made is between directly succeeding actions or set of actions keeping some order, but allowing other in-between actions. The answer to this question will definitely lead to the different analysis, challenges and solutions.

Another challenge lies within the temporal aspect of the processes. Each action can have its start time and completion time, therefore allowing for some temporal relations in between. Sometimes the actions can be performed in "parallel", in other cases the order may be important. A challenge is how to deal with parallel actions or distinguish between such situations.

Even if we consider processes with no parallel actions there are still many challenges regarding the sets of subsequent actions. One of the biggest is how we define the similarity between to sequences or subsequences. Traditionally in process mining Levenshtein distance is used [5]. It is a string metric for measuring the difference between two sequences and is computed as the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. However, even if we decide to use the Levenshtein distance or any version of string edit distance, there are some open questions that remains open. Let us consider here a few examples.

Let us assume we are interested in subsequence "A B C". It is easy to check if this sequence is present. However, we should ask how to proceed if the sequence is only partially present, for instance in cases "A B D" or "A B D C". Another issue may be raised if a sequence in question is present more than once, as in the case of "**A B C** F **A B C** D". Should this double (or in general case multiple) occurrence of the sequence under consideration influence the membership degree?

It may seem obvious, that the assumed similarity measure should be equal to 1 when the sequence in question is present in the trace. Similarly, zero

SR #	Date	Status	Sub Status	Funct. Div	Org. line	ST	Im-pact	Prod	C1	Owner C.	Owner Name
1-364	2010-03-31 15:59:42	Accepted	In Progress	A2_4	A2	V30	Med	582	fr	France	Fred
1-364	2010-03-31 16:00:56	Accepted	In Progress	A2_4	A2	V30	Med	582	fr	France	Fred
1-364	2010-03-31 16:45:48	Queued	Awaiting	A2_5	A2	V5	Med	582	fr	France	Fred
1-364	2010-04-06 15:44:07	Accepted	In Progress	A2_5	A2	V5	Med	582	fr	France	Anne
1-364	2010-04-06 15:44:38	Queued	Awaiting	A2_4	A2	V30	Med	582	fr	France	Anne
1-364	2010-04-06 15:44:47	Accepted	In Progress	A2_5	A2	V13	Med	582	fr	France	Anne
1-364	2010-04-06 15:44:51	Completed	Resolved	A2_5	A2	V13	Med	582	fr	France	Anne
1-364	2010-04-06 15:45:07	Queued	Awaiting	A2_4	A2	V30	Med	582	fr	France	Anne
1-364	2010-04-08 11:52:23	Accepted	In Progress	A2_4	A2	V30	Med	582	fr	France	Eric
1-364	2010-04-08 11:53:35	Queued	Awaiting	A2_5	A2	V5	Med	582	fr	France	Eric
1-364	2010-04-20 10:07:11	Accepted	In Progress	A2_5	A2	V5	Med	582	fr	France	Anne
1-364	2010-04-20 10:07:19	Accepted	Assigned	A2_5	A2	V5	Med	582	fr	France	Anne

Table 1: An excerpt from Volvo IT incident management system log [35].

similarity is observed when none of the actions of interest is present in the trace. Yet, how do we normalize those similarity measures? Should we incorporate only the length of the sequence, or also the positions? For instance, are the pair of sequences "A B" and "A C" similar to the same degree as the pair "A B C D" and "A B F G"? Note that both have the half of actions the same, and half of actions different. Do the indices on which the sequences differ matter? Consider the pairs "A B C D" – "A B F G" and "A B C D" – "A G C F". Are they similar to the same degree or not?

Summarizing the sequences of actions will require answering such questions. We believe there are no good or bad decisions, but they will have different consequences for the interpretation of the summary.

4.3. New protoforms

In classical Yager's approach [41] there were two types of protoforms: simple one (Q y's are P) and extended one (Q R y's are P), where Q is the quantifier, y's are the objects that are summarized, P is the summarizer and R is the qualifier.

The question is if the process data require new type of prototypes, or can the already proposed prototypes be adapted to the new data. The new prototypes should be found useful and meaningful by the potential users and domain experts.

There is also the question of what are useful features that can serve as summarizers, but this question depends a lot on the application area. If we consider the Volvo IT event log [35], one may be interested in a summary such as; "many cases start with the sequence *In Progress, Awaiting Assignment, In Progress*". In case of a healthcare related data, we could consider summary as "most patients had a long waiting time". Proposed methods should take into account possible differences amongst different domains.

4.4. Different perspectives

Traditionally, three different perspectives [34] are distinguished in process mining:

- process perspective focused on control flow, ordering of activities,
- case perspective, focusing on properties of cases, and
- organizational perspective, focusing on the performers, which are involved and how they are related.

Process mining treats these perspectives separately. However we believe combining them together may bring some additional knowledge and benefit, providing some additional knowledge and quality. An example in which we are combining the case perspective with the organizational perspective from Volvo IT dataset [35] is: "in many cases when impact was high the problem was transfered from unit A2_4 to C2". Also, interesting summaries may be obtained for the health-care, *e.g.,*:"most cases where the patient was a small child, had a long throughput time".

4.5. Causality

Processes have much more to offer than sequence analysis. Finding relationships between different features of processes may be interesting and beneficial for the user. However, the real added value is in finding the causality relationships between the actions or process features.

We may question ourself whether we can discover that an action or event is a trigger for another action, and if yes, how this can be achieved. The challenge may be expressed as how to determine that there is indeed a causality relationship, not only a pure coincidence. Discovering those causality relationships may be a first step for triggering new insights into the interaction between process execution and process outcomes. For example, in the health domain, one could identify whether following one care path over another leads to decreased number of complications, indicating an improved care delivery.

4.6. Usuality

While talking with domain experts regarding processes, they often use the word *usually*. Linguistic summaries are so far expressed using possibilistic modality. We are not aware about much work on different modalities for linguistic summaries. Zadeh in [43, 44] provided an outline for the theory of usuality, in which he proposed that "usually (X is F)" is equivalent to "most $x \in X$ are F". However this topic requires further investigation. There is also a need for defining a more general approach for other usuality related quantifiers like "seldomly".

Another related question concerns typicality. It may be useful and interesting to investigate what the typical trace of a process is, or typical subsequence. There is an open question now to match this concept with linguistic summaries.

4.7. Generation of the summaries

Last but not least, the common challenge for all types of linguistic summaries is the generation process (cf. Kacprzyk and Wilbik [15], Pilarski [26], Castillo-Ortega *et al.* [9]). There are two main issues: efficiency and completeness. We wish to obtain the summaries quite fast. We also wish to obtain all useful summaries, so that we can build a whole picture, but not too many, in order to be comprehensible by the humans.

It is relatively easy to evaluate the validity (truth) of summaries. There are many approaches and papers on the this topic. A nice overview of evaluating quantified sentences can be found in Delgado et al. [10]. However, the truth value is not enough, and therefore several other criteria were introduced [18, 19], like specificity, appropriateness, and informativeness. Yet, the challenge still remains whether those summaries are useful for the user and whether the description is complete.

Even a simple event log may have many attributes. For example, Volvo IT log [35] has 12 attributes for each entry. This creates many possibilities for many possible summarizers, such as throughput time, length of certain sequence of actions, involvement of certain peoples or resources, product types, countries, etc. Therefore there are many possibilities of linguistic summaries that can be checked. This number gets even bigger if we allow combining those attributes using conjunctions.

In the Volvo example, for instance, summarizing cases like those that refer to "a certain type of the product, and people from France were involved in solving the problem" leads to an explosion of the number of combinations.

4.8. Bottleneck detection

The purpose of the process analysis is to understand and improve the underlying processes. Automated detection of the problems and bottlenecks is still a challenge. Will the linguistic summarizes have the capability of bottleneck detection? Will it be possible to find the reasons that causes certain problems, so that some actions can be taken? The challenge is to find methods, frameworks that will allow to answer the above two questions positively.

5. A research agenda

In this paper we presented eight challenges for linguistic summaries of processes. Challenge 1 – the event logs is also a challenge for the process mining community in general. It would be worth to combine forces and learn from each other.

Challenge of creating new protoforms and dealing with sequences are specific to the topic of linguistic summaries. Although we indicated one protoform possibility, we believe there are many more options, depending on the users' questions. Those two challenges are quite important and they should be addressed as first. Challenge 4, looking at different perspectives, may be quite intuitively combined with the previous 2 challenges, and may bring additional value to the users.

Finding causal relations was mentioned as the next. This topic may be very interesting, and useful but also complicated to find a solution. Therefore we think sufficient progress must be made in other challenges before this one can be addressed.

Next two of the named problems, namely usuality and an efficient way of generating the summaries are not process specific. They are present also in linguistic summarization from simple attribute–value of data bases.

The last of the mentioned challenges is the bottleneck detection. We believe that it may use the results of all previous challenges mentioned here, and can be seen as the long term goal for the linguistic summaries of processes.

6. Concluding remarks

In this paper, we considered the challenges for obtaining linguistic summaries from process data. So far, various linguistic summarization techniques have been proposed for data sets consisting of attribute-value pairs. However, process data are characterized by relational structure in which explicit relations amongst data elements exist and are typically represented in a graph. Therefore, addressing those differences that structured data have compared to attribute-value pairs, we acknowledged the need of investigating this topic.

We believe that linguistic summaries of process data will be a valuable tool for process analysts. Due to complexity and diversity of observed behaviors, traditional tools may generate visual models with spaghetti-like pathway patterns that are difficult to comprehend for humans. However linguistic summaries, use a different communication mechanism, natural language, which is a natural way of communication for humans. Linguistic summaries have the advantage that they may combine different perspectives easily, like control-flow, organizational and case perspectives, which can provide new insight into the existing processes.

References

[1] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M. Aud. Linguistic summarization of video for fall detection using voxel person and fuzzy logic. *Computer Vision and Image Understanding*, 1(113):80–89, 2009.

[2] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M. Aud. Modeling human activity from voxel person using fuzzy logic. *IEEE Transactions on Fuzzy Systems*, 1(17):39–49, 2009.

[3] D. Anderson, R. H. Luke, E. Stone, and J. M. Keller. Segmentation and linguistic summarization of voxel environments using stereo vision and genetic algorithms. In *Proceedings IEEE International Conference on Fuzzy Systems, World Congress on Computational Intelligence*, pages 2756–2763, 2010.

[4] P. Bosc, D. Dubois, O. Pivet, H. Prade, and M. D. Calmes. Fuzzy summarization of data using fuzzy cardinalities. In *Proceedings of IPMU 2002 Conference*, pages 1553–1559, 2002.

[5] R. J. C. Bose and W. M. van der Aalst. Context aware trace clustering towards improving process mining results. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009*, pages 401–412, 2009.

[6] B. Bouchon-Meunier and G. Moyse. Fuzzy linguistic summaries: where are we, where can we go? In *IEEE Conference on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, pages 317–324, 2012.

[7] R. Castillo-Ortega, N. Marín, and D. Sánchez. Time series comparison using linguistic fuzzy techniques. In E. Hüllermeier, R. Kruse, and F. Hoffmann, editors, *Computational Intelligence for Knowledge-Based Systems Design, Proceedings of the 13th International Conference on Information Processing and Management of Uncertainty, IPMU 2010*, pages 330–339. Springer, 2010.

[8] R. Castillo-Ortega, N. Marín, and D. Sánchez. Linguistic local change comparison of time series. In *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, pages 2909–2915, June 2011.

[9] R. Castillo-Ortega, N. Marín, D. Sánchez, and A. G. Tettamanzi. Linguistic summarization of time series data using genetic algorithms. In *EUSFLAT*, pages 416–423. Atlantis Press, 2011.

[10] M. Delgado, M. D. Ruiz, D. Sanchez, and M. A. Vila. Fuzzy quantification: a state of the art. *Fuzzy Sets and Systems*, 242:1 – 30, 2014.

[11] D. Dubois and H. Prade. Gradual rules in approximate reasoning. *Information Sciences*, 61:103–122, 1992.

[12] D. Dubois, H. Prade, and E. Rannou. User-driven summarization of data based on gradual rules. In *Proceedings of the Sixth IEEE International Conference on Fuzzy Systems*, volume 2, pages 839–844, 1997.

[13] M. Dumas, W. M. van der Aalst, and A. H. ter Hofstede. *Process-Aware Information Systems: Bridging People and Software Through Process Technology*. Wiley, 2005.

[14] J. Kacprzyk and P. Strykowski. Linguistic summaries of sales data at a computer retailer: a case study. In *Proceedings of IFSA'99*, volume 1, pages 29–33, 1999.

[15] J. Kacprzyk and A. Wilbik. Towards an efficient generation of linguistic summaries of time series using a degree of focus. In *Proceedings of the 28th North American Fuzzy Information Processing Society Annual Conference – NAFIPS 2009*, 2009.

[16] 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.

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

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

[19] J. Kacprzyk, R. R. Yager, and S. Zadrożny. Fuzzy linguistic summaries of databases for an

efficient business data analysis and decision support. In W. Abramowicz and J. Żurada, editors, *Knowledge Discovery for Business Information Systems*, pages 129–152. Kluwer, Boston, 2001.

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

[21] J. Kacprzyk and S. Zadrozny. Supporting decision making via verbalization of data analysis results using linguistic data summaries. In E. Rakus-Andersson, R. R. Yager, N. Ichalkaranje, and L. C. Jain, editors, *Recent Advances in Decision Making*, volume 222, pages 121–143. Springer Berlin Heidelberg, 2009.

[22] U. Kaymak, R. Mans, T. van de Steeg, and M. Dierks. On process mining in health care. In *Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2012)*, pages 1859–1864, October 2012.

[23] D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[24] G. Moyse, M.-J. Lesot, and B. Bouchon-Meunier. Linguistic summaries for periodicity detection based on mathematical morphology. In *2013 IEEE Symposium on Foundations of Computational Intelligence (FOCI)*, pages 106–113, 2013.

[25] G. Moyse, M.-J. Lesot, and B. Bouchon-Meunier. Mathematical morphology tools to evaluate periodic linguistic summaries. In *Proceedings of FQAS 2013*, pages 257–268, 2013.

[26] D. Pilarski. Linguistic summarization of databases with quantirius: a reduction algorithm for generated summaries. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 18(3):305–331, 2010.

[27] M. R. Rajati and J. M. Mendel. Advanced computing with words using syllogistic reasoning and arithmetic operations on linguistic belief structures. In *FUZZ-IEEE 2013, IEEE International Conference on Fuzzy Systems, Hyderabad, India, 7-10 July, 2013, Proceedings.*, pages 1–8, 2013.

[28] G. Raschia and N. Mouaddib. Using fuzzy labels as background knowledge for linguistic summarization of databases. In *The 10th IEEE International Conference on Fuzzy Systems 2001*, volume 3, pages 1372–1375, 2001.

[29] G. Raschia and N. Mouaddib. SAINTETIQ: a fuzzy set-based approach to database summarization. *Fuzzy Sets and Systems*, 129:137–162, 2002.

[30] D. Rasmussen and R. R. Yager. Finding fuzzy and gradual functional dependencies with SummarySQL. *Fuzzy Sets and Systems*, 106:131–142, 1999.

[31] M. Ros, M. Pegalajar, M. Delgado, A. Vila, D. T. Anderson, J. M. Keller, and M. Popescu. Linguistic summarization of long-term trends for understanding change in human behavior. In *Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2011*, pages 2080–2087, 2011.

[32] P. S. Szczepaniak and J. Ochelska. Linguistic summaries of standardized documents. In M. Last, P. S. Szczepaniak, Z. Volkovich, and A. Kandel, editors, *Advances in Web Intelligence and Data Mining*, pages 221–232. Springer Berlin Heidelberg, 2006.

[33] W. M. van der Aalst. *Process Mining – Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011.

[34] B. F. van Dongen. *Process Mining and Verification*. PhD thesis, Technische Universiteit Eindhoven, 2007.

[35] Ward Steeman. Volvo IT Belgium, closed cases event log, `doi:10.4121/c2c3b154-ab26-4b31-a0e8-8f2350ddac11`.

[36] A. Wilbik and J. Kacprzyk. Temporal sequence related protoforms in linguistic summarization of time series. In *proceedings of the World Conference on Soft Computing 2011*, 2011.

[37] A. Wilbik and U. Kaymak. Gradual linguistic summaries. In *Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2014,Part II*, pages 405–413, 2014.

[38] A. Wilbik, J. M. Keller, and G. L. Alexander. Linguistic summarization of sensor data for eldercare. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC 2011)*, pages 2595–2599, 2011.

[39] A. Wilbik, J. M. Keller, and J. C. Bezdek. Linguistic prototypes for data from eldercare residents. *IEEE Transactions on Fuzzy Systems*, 22(1):110–123, 2014.

[40] D. Wu and J. M. Mendel. Linguistic summarization using if – then rules and interval type-2 fuzzy sets. *IEEE Transactions on Fuzzy Systems*, 19(1):136–151, 2011.

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

[42] R. R. Yager, K. M. Ford, and A. J. Cañas. An approach to the linguistic summarization of data. In B. Bouchon-Meunier, R. R. Yager, and L. A. Zadeh, editors, *Uncertainty in Knowledge Bases, 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU '90, Paris, France, July 2-6, 1990, Proceedings*, pages 456–468. Springer, 1990.

[43] L. Zadeh. Outline of a theory of usuality based on fuzzy logic. In A. Jones, A. Kaufmann, and H.-J. Zimmermann, editors, *Fuzzy Sets Theory*

and Applications, pages 79–97. NATO ASI Series, 1986.

[44] L. A. Zadeh. Fuzzy sets, fuzzy logic, and fuzzy systems. chapter Outline of a Theory of Usuality Based on Fuzzy Logic, pages 694–712. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1996.

[45] S. Zadrozny and J. Kacprzyk. Summarizing the contents of web server logs: A fuzzy linguistic approach. In *Proceedings of FUZZ-IEEE 2007*, 2007.