Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems ••• (••••) •••–•••

FUZZY sets and systems

www.elsevier.com/locate/fss

# A soft computing approach to big data summarization

Grégory Smits [a,*], Olivier Pivert [a], Ronald R. Yager [b], Pierre Nerzic [a]

[a] *IRISA - University of Rennes 1, UMR 6074, Lannion, France*
[b] *Machine Intelligence Institute - IONA College, New Rochelle, NY, USA*

Received 27 June 2016; received in revised form 16 February 2018; accepted 28 February 2018

**Abstract**

The added value of a dataset lies in the knowledge a domain expert can extract from it. Considering the continuously increasing volume and velocity of these datasets, efficient tools have to be defined to generate meaningful, condensed and human-interpretable representations of big datasets. In the proposed approach, soft computing techniques are used to define an interface between the numerical and categorical space of data definition and the linguistic space of human reasoning. Based on the expert's own vocabulary about the data, a personal summary composed of linguistic terms is efficiently generated and graphically displayed as a term cloud offering a synthetic view of the data properties. Using dedicated indexing strategies linking data and their subjective linguistic rewritings, exploration functionalities are provided on top of the summary to let the user browse the data. Experimentations confirm that the space change operates in linear time wrt. the size of the dataset making the approach tractable on large scale data.


*Keywords:* Data personalisation; Linguistic summaries; Soft computing; Knowledge extraction; Visualization; Specificity measure

## 1. Introduction

Data analysis is a crucial task at the center of many professional activities and now constitutes a support for decision making, communicating and reporting. Considering the continuously increasing volume and velocity of these datasets, domain experts (as insurers, data journalists, communication managers, decision makers, etc.), who are not, most of the time, data or computer scientists, need efficient tools that help them turn data into useful knowledge. This explains the recent growing interest for so-called Agile Business Intelligence (ABI) systems that reconsider classical data integration processes to favor pragmatic approaches that make domain experts self-reliant in the analysis of raw data.

A dataset generally consists of a large collection of items described by numerical and categorical attributes. A way to assist experts in their fastidious task of data-to-knowledge translation is to define efficient strategies that generate meaningful, condensed and human-interpretable representations of the data. To be very useful, such representations should give an insight into the data properties and make it easy for the domain expert to identify the most representative

* Corresponding author.
*E-mail addresses:* gregory.smits@irisa.fr (G. Smits), olivier.pivert@irisa.fr (O. Pivert), yager@panix.com (R.R. Yager), pierre.nerzic@irisa.fr (P. Nerzic).

https://doi.org/10.1016/j.fss.2018.02.017
0165-0114/

properties of the dataset. In this sense, when a dataset is so large that it cannot be easily perused and analyzed by the user, data summarization is of a particular interest to obtain a big picture of the data distribution on the different dimensions. Such a summary should also offer exploration functionalities to let the expert interactively browse the dataset from its summary and discover interesting properties possessed by different data subsets.

The approach proposed in this paper falls in with the original essence of soft computing as it aims to create an interface between the numerical/categorical space of data definition and the symbolic space of human reasoning. Based on an expert vocabulary materialized by fuzzy partitions and linguistic variables, efficient algorithms are provided to rewrite the data according to subjective linguistic terms taken from the expert's own vocabulary. To help the expert identify the linguistic terms that best describe a data subset, we provide a specificity-like measure that quantifies the representativity of a term wrt. the concerned data subset. The result of the rewriting process is materialized by a so-called rewriting vector that tells us about the distribution of the data according to the different linguistic terms of the expert's vocabulary. The fact that the numerical and categorical properties possessed by items of the dataset are all rewritten into linguistic terms makes it possible to envisage unified and global views of the data. Contrary to most of the existing approaches to ABI that are indeed only able to generate two or three-dimensional graphical views of the data, using e.g. histograms, charts or color maps, novel graphical representations of the linguistic rewriting vectors may be defined to provide the expert with a complete and concise view of the data, on several dimensions, using linguistic terms instead of numbers.

The main contributions of this paper concern the proposal of:

- efficient algorithms for data rewriting using linguistic terms taken from the expert's vocabulary,
- a measure that quantifies the informativeness of the linguistic terms wrt. a data subset,
- dedicated storage and indexing strategies,
- a condensed and human-interpretable graphical view of the data using linguistic terms.

The rest of the paper is organised as follows. After having compared the context and objective of the proposed approach wrt. to existing works (Sec. 2), Sec. 3 introduces some preliminary notions to this approach. Then, Sec. 4 details the rewriting process as well as storage strategies to associate items with their respective linguistic rewritings. Section 5 shows how the linguistic rewriting of a dataset may be displayed as a term cloud. Before concluding, Sec. 6 gives the results of an experimentation on a large dataset that confirm the efficiency of the proposed rewriting and visualization strategy.

### *1.1. Approach overview*

Fig. 1 depicts the different steps of the approach proposed in this work. Faced with a raw dataset to analyze, the expert is first solicited to define, on attributes of interest, linguistic terms that will form his/her vocabulary and that will then be used to handle the data as well as the extracted knowledge. Each item from the dataset to analyze is first linguistically rewritten according to the user's vocabulary, these rewritings being stored in a database. A global rewriting vector is also computed for the whole dataset to represent the distribution of the items wrt. the different linguistic terms of the vocabulary. This dataset rewriting vector, that is very small compared to the size of the dataset, is then graphically rendered to the user, as a so-called term cloud, so as to provide a linguistic and subjective summary of the dataset. A term cloud offers a graphical view of the data distribution wrt. the terms of the user's vocabulary. Thanks to the fact that one stores the items and their associated rewritings in a database, efficient and intuitive data exploration functionalities are offered to the user to let him/her browse the data by successively selecting terms from the cloud. The cloud shown in Fig. 1 contains the terms that describe the items satisfying a previously selected property, here 'price is acceptable'. A measure is used to quantify the informativeness of each term appearing in a cloud so as to display the most informative terms in the eye-catching zone of the view, its center.

## 2. Related work

ABI is a recent issue raised by the need for domain experts to quickly analyze a large number of raw datasets and decide if their integration in the corporate datamart is useful or not. Driven by user needs, this issue has first been addressed by the main data management companies (e.g. Amazon with QuickSight, Pentaho's Agile BI, Microsoft

Please cite this article in press as: G. Smits et al., A soft computing approach to big data summarization, Fuzzy Sets Syst. (2018), https://doi.org/10.1016/j.fss.2018.02.017