International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 9, No.5 (Sep-2020)

http://dx.doi.org/10.12785/ijcds/090505

Data Summarization for Sensor Data Management: Towards Computational-Intelligence-Based Approaches

Boulanouar Khedidja[1,2], Hadjali Allel[1] and Lagha Mohand[2]

[1] LIAS/ENSMA, 1 Avenue Clement Ader, 86960 Futuroscope Cedex, France
[2]University Blida1, Laboratory of Aeronautical Sciences, Institute of Aeronautics and Spatial Studies, Algeria

Received 2 Nov.2019, Revised 12 May 2020, Accepted 1 Aug. 2020, Published 1 Sep. 2020

Abstract: The rapid and significant increase in the amount of sensor data to be processed requires the use of techniques to reduce the size of data in order to efficiently extract the relevant knowledge. In this paper, we present two approaches used to derive data summaries. The first one relies on linguistic quantifiers in the sense of Yager. The second one leverages the notion of the typical value of a data set. Then, we present the implementation of these two methods with some experiments conducted on different databases (real-flight data collected from the ADSB project and real data for smart city collected from neOCampus project). Finally, a comparative study is discussed to show the best approach w.r.t. execution time.

Keywords: Data summarization, Fuzzy Logic, Linguistic Quantifiers, Typical Value.

1. INTRODUCTION

Thanks to the big progress of sensor technologies, a large number of applications have emerged (smart environments, mobility and intelligent tourism, public safety and environmental monitoring, etc.) where huge quantities of data can be continuously and quickly generated. These data represent the raw material for decision making. However, the huge volume of such data makes their processing very expensive in terms of energy and computing time. Therefore, new methods to large sensor data management are needed to reduce this cost of processing, especially, in the application domains where approximate answers are sufficient for decision-making.

Data reduction is an interesting and promising way to achieve the above goal. Its main principle is to re-write the original data in a compact and concise form to leverage this form for the decision-making purpose. The notable advantage to consider a data compact form is the fact that it allows answering the user query with an acceptable cost in terms of execution time and energy consumption. This is particularly highly desirable in real-life applications where the real-time aspect of answers is critical, and their approximate features bring enough information to be acceptable. Summarization constitutes an appropriate technique to perform a reduction of large amounts of data.

Summarization [1] is a process of creating a concise, yet informative, version of the original data. The terms concise and informative are quite generic and depend on application domains. Summarization has been extensively studied in many domains including text analysis, network traffic monitoring, financial domain, health sector and many others. Since summarization uses the semantic content of the data, it has been proven to be a useful and effective data analysis technique for interpreting large-scale datasets. For example, summarization is an important step to expedite knowledge discovery from data tasks (which are time-consuming) by intelligently reducing the size of processed data.

One can distinguish several data summarization methods. The most known among them Sampling, Sketching, Histogramming, Wavelets, Aggregates, etc. They rely on statistical techniques that usually describe the statistical characteristics of attributes. Despite the fact that, the statistical summarization is easy to implement for extracting knowledge, it provides limited knowledge [2]. In particular, they suffer from some non-negligible limitations in the Big data context, namely: (i) lack of representativeness of the original data; (ii) low recovery of the complex and diverse decision-makers' needs; (iii) the summary structures built are hardly intelligible. In this paper, we are interested in a new generation of summary structures that are borrowed from the computational intelligence field [3-9]. In particular, two data summarization methods are discussed: non-classical quantifiers-based and typicality-based method. The interest of such methods is twofold: the summaries produced are intelligible and allow describing the original data at different levels of abstraction. We show in detail how such methods can be applied also we study the properties of the produced summaries.

The remainder of the paper is structured as follows. In section 2, we give an overview of data summarization. We

E-mail: khedidja.boulanouar@ensma.fr,allel.hadjali@ensma.fr,mlagha-aerospatiale@univ-blida.dz

present then the linguistic-quantifiers-based summaries of data in section 3. The notion of typical value and the typicality-based summaries are discussed in section 4. Section 5 provides and analyses the results of our experiments. Section 6 concludes the paper and provides some future directions.

2. Related Work

The increasing volume of data makes their processing difficult and expensive in many cases. It becomes even impossible to store all the data that would sometimes be desirable to keep as it is generated at a fast pace or in large quantities.

This situation has called for the creation of a new vision of data reduction, which has been the subject of growing interest from different communities (databases, data analysis, data mining, etc.) in industrial and academic fields.

Several studies have been proposed in the literature [10-13]. In [13], the author describes the different techniques of structured and unstructured data reduction, for example, the case of sampling where the objective is to provide information on a large population of data from a representative sample extracted from it. Various categories of sampling are cited, the most popular are: simple random sampling, stratified random sampling, cluster random sampling.

Another way to summarize data is to use histograms. This structure is often used to summarize qualitative or quantitative data. A histogram separates the population into a set of groups or classes according to attributes.

The notion of sketch was introduced for the first time by [14]. It is a very compact data flow summary that is used to estimate the response to certain queries on all data. This technique aims to randomly project each element into space using hash functions and to keep only the most relevant components, thus saving space while preserving most of the information.

Unfortunately, each technique has some important drawbacks that will limit their exploitation; they are still far from being able to reflect a real human perception, due to a little useless of natural language. However, it is worth looking at some nature-inspired solutions, which mainly rely on the way the human brain handles continuous incoming sensory data by harnessing contextual characteristics of learning processes (see, e.g., [15]).

It should be noted that these summary structures capture certain properties and statistics that are very useful and very interesting for decision-makers, but they suffer from certain limitations as mentioned in the introduction. For this purpose, a new generation of summary structures from the field of computational intelligence has found its place in the literature; among these techniques, we can find quantification linguistic summary and the concepts of typicality. In [4], Yager gave birth to these two notions, then He introduced the linguistic summary paradigm in [5][7][8]. Two forms of the summary are studied “Q y are S” and “Q R y are S”. The quality of the summary produced is also discussed. Another trend in database summarization is proposed by Dubois et al. [9]. The authors use the concept of typical value, more precisely, fuzzy typical value in order to generate summaries from a dataset.

3. Linguistic-Quantifiers-Based Summary

For end-users, it is easier to understand a natural language statement such as *"most of temperatures are high"* than to understand statistical characteristics of attributes such as median, mean value, standard deviation and so on. Such statement provides a concise and intelligible description of the semantics content of data which is called *linguistic summary* in the literature.

In the following, we revisit the approach of Yager [4] to summarize datasets of interest with linguistic quantifiers. First, let us briefly present the main idea of this approach which constitutes the basic steps for our derivation of linguistic summaries. Assume that

1. V is an attribute or a quality that can take values in the set $X = \{x_1, x_2, \dots, x_n\}$. For example, V could be temperature, pressure or any other quality.
2. $Y = \{y_1, y_2, \dots, y_n\}$ is a set of objects (or records) that manifest the quality V; hence $V(y_i)$ is the value of V for the object y_i.
3. $D = \{V(y_1), V(y_2), \dots, V(y_n)\}$ the collection of data that we want to summarize.

The basic structure of linguistic summary is "Q entities are/have S"; so, one can say that the summary consists of three items (figure 1 shows the different summary components):

1. A summarizer S (e.g., high).
2. A quantity in agreement Q (e.g., most).
3. A measure of validity (or truth) of summary T (e.g., 0.8).

We can then write for instance *"T (most of temperatures are high) = 0.8)"*. The truth T may be meant in a more general sense, e.g. as validity of, even more generally, as some quality or goodness of a linguistic summary.

Now, let us consider a set of data D. We can hypothesize any suitable summarizer S and any quantity in agreement Q, and the assumed truth measure will indicate the verity or truth of the statement that Q data items satisfy the summarizer S [16].

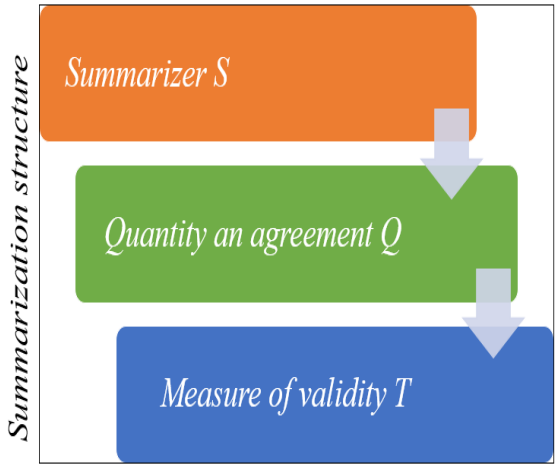

Figure 1. Linguistic quantifiers-based approach

Let us start to explain the two former elements and then address the issue of how to calculate the degree of truth.

FORM OF SUMMARIZER

Since humans basically depend on natural languages as the main means of communication, we suppose that the summarizer S is a linguistic expression semantically modeled by a fuzzy set [16]. For instance, in *"most of the Temperatures are high"*, a summarizer like ''high'' would be represented as a fuzzy set-in the universe of discourse containing possible values of the temperature.

FORM OF QUANTITY IN AGREEMENT:

The quantity in agreement Q also called the fuzzy quantifier in [17] is a linguistic term that represents the number of elements from the data which satisfy the summary. Two types of quantifiers can be employed:

- Absolute: represents a fuzzy set that takes values in the set of real non-negative numbers. For instance, "about 5", "more or less 100", "several", etc.
- Relative: represents a fuzzy set of the unit interval. For instance, "a few", "more or less a half", "most", "almost" all, etc.

TRUTH COMPUTATION

The calculation of the truth (or, more generally, validity or quality) of the basic type of a linguistic summary considered in this section is equivalent to the calculation of the truth-value (from the unit interval) of a linguistically quantified statement (e.g., "*most of the temperature are high*"). This may be done by two most relevant techniques using either Zadeh calculus of linguistically quantified statements [17] or Yager OWA operators [18].

Using the protoform proposed by Yager [4] "Q y's are S" where Q is the fuzzy quantifier, y is the set of data and S is the summarizer. The truth degree is obtained by using the following formula:

$$T = \mu_Q\left[\frac{\sum_{i=0}^{n} \mu_S(y_i)}{n}\right] \quad (1)$$

where n represents a scalar cardinality (the number of objects of the dataset).

In [16] the authors introduced another form of quantified proposition represented by (Q R y's are S) where R is a linguistic expression indicates the qualifier. In this case, the truth degree can be computed by (2)

$$T = \mu_Q\left[\frac{\sum_{i=0}^{n} \min(\mu_S(y_i).\mu_R(y_i))}{\sum_{i=0}^{n} \mu_R(y_i)}\right] \quad (2)$$

An Illustrative Example:

Assume we have a collection of data describing the Temperature:

$$D = \{15,21,24,23,22,23,25,29,30,17\}$$

Assume a proposed summary of this data collection is: summarizer ="high", quantity in agreement = "most". These two concepts are defined by the user as two fuzzy subsets, the summarizer membership function can be defined as follows:

$$S(x) = \begin{cases} 0 & x < 20 \\ \frac{x}{5} - 4 & 20 \le x \le 25 \\ 1 & x > 25 \end{cases}$$

for each object $d_i \in D$, we calculate $S(d_i)$, the degree of summarizer satisfaction as represented in table 1:

TABLE I. SUMMARIZER SATISFACTION DEGREE

d_i	$S(d_i)$	d_i	$S(d_i)$
15	0	23	0.6
21	0.2	25	1
24	0.8	29	1
23	0.6	30	1
22	0.4	17	0

The quantifier "most" can be also defined as follows.

$$\mu(x) = \begin{cases} 1 & x \ge 0.7 \\ 5x - 2.5 & 0.5 \le x < 0.7 \\ 0 & x < 0.5 \end{cases} \quad (3)$$

Now, let $r = \frac{1}{n} * \sum_{i=1}^{n} S(d_i) = 0.56$, the portion of D that satisfies S. One can easily check that

$$T = Q(r) = Q(0.56) = 0.3$$

Then, the validity of the summary "*most of D are high*" is 0.3.

SOME OTHER VALIDITY CRITERIA

According to [19], the basic validity criterion, i.e. the truth of a linguistically quantified is certainly the most important. However, it does not grasp all aspects of a linguistic summary. The authors attempted to devise other validity criteria; they proposed new qualities of linguistic database summaries such

- a degree of imprecision (fuzziness),
- a degree of covering,
- a degree of appropriateness,
- a length of a summary.

4. TYPICAL VALUE

The concept of typicality has been studied using several methods for the data summarization purpose. According to [20, 21], the typical value

- could be defined as a value that is the same or very similar to most of the observations in the collection of data we are trying to typify;
- should be the most frequent values in the dataset as well.

In [20], a way to summarize the possible values in a scalar way is proposed. This approach shows that the average may be not completely satisfactory from the user's point of view. They began by defining the typicality in general, what are its weaknesses, and why it's better to find another way to measure the typicality of a set against the traditional methods. Then, they described how one can calculate the typical value and find the typical measure based on the compatibility index and the specificity index. The two indexes must belong to the unit interval. The first one refers to the degree of satisfaction when a proposed fuzzy subset A satisfies the concepts to be a typical value for a collection of data D. The specificity index represents the requirement of the narrowness of the typical value requires the introduction of the concept of specificity.

In [9] the authors found that this method has some limits because the founded interval could be larger than the beginning set. According to these critics, they presented another approach to find I* the best optimal interval based on the density of classification $f(x_i, l)$, the extraction of $l^*$the optimal typical step between two range of the interval I^* from $f(x_i, l)$ and the extraction of the maximum of the best proportion.

They also defined the interval $I(x_i, l) = [x_i, x_i + l]$ where l is the step from [0, L] (L is the maximum value of the data collection) and the cost function $f(x_i, l)$ which is the probability to select a value from the interval $I(x_i, l)$.

$$f(x_i, l) = \frac{|I(x_i, l)|}{n} \quad (4)$$

For given value l the "most Typical" interval $I^*(x_i, l)$ is that maximizes the cost function [9].

An Illustrative Example:

To better explain the approach, we present in table 2 an illustrative example in order to show how we can obtain the optimal interval.

Let $D = \{0,3,4,5,6,7,8,9,12,23\}$ be a collection of observations. Let $X = \{1,1,1,4,7,5,3,5,2,1\}$stands for the occurrences of each $d_i \in D$.

We have L = 23: the maximum value of the collection. We chose the step $l \in [1, L]$, we can note an interval I by: $I(x_i, l) = [x_i, x_i + l] \subset X$ and we use the cost function defined in equation (4).

In this example, we start by computing $f(x_i, l)$ for each interval I and step l. An induced value is then the interval with the maximum values of $f(x_i, l)$, for each l.

To obtain the typical value of the beginning set D, we have to compute $f(x_i, l) - \frac{l}{L}$ and consider the interval I* as the best definition of values.

TABLE II. AN ILLUSTRATIVE EXAMPLE OF THE TYPICAL VALUE

l	I^*	$f(xi, l)$	l/L	$f^*(xi, l) - l/L$
1	[6,7]	12/30	1/23	0.3565
2	[5,7]	16/30	2/23	0.4460
3	[6,9]	20/30	3/23	0.5360
4	[5,9]	24/30	4/23	0.6266
5	[4,9]	25/30	5/23	0.6159

By simple interpolation on the curve, we obtain $l^* = 4$ with the best cost function $f(x_i, l) = 24/30$. We can deduce next the interval [5, 9] as the optimal typical value in D. We follow algorithm 1 which is a formalization of the approach of Dubois and Prade, introduced in [9] in order to extract the optimal typical value.

Algorithm 1 Dubois and Prade Algorithm

```
for l = 1 ; l < L ; l = l + 1 do
  for x_i = x_0; x_i ≤ n; x_i + l do
    f(x_i, l) = Card(I(x_i, l))/n
    if f(x_i, l) ≥ max1 then
      max1 ← f(x_i, l)
      I ← I(x_i, l)
    end if
  end for
  if f(x_i, l) > max2 then
    max2 ← f(x_i, l) − l/L
    I* = I
  end if
end for
```

5. SUMMARIES EVALUATION AND EXPERIMENTATION

The increasing volume of data has therefore brought about the need for a new generation of computational techniques and tools for extracting useful knowledge. These techniques and tools have led to the emerging field

of knowledge discovery/data mining, as an increasingly important research area[22, 23].

The first main goal of the summary is to extract knowledge from a large set of data and represent it in a condensed form and sensitive manner. The second goal is to improve response time.

The proposed procedures of summarization were implemented on a real data, the first database is a static database representing flight data recorders; these data are gathered from sample flight data project which is Automatic Dependent Surveillance-Broadcast (ADSB) project. The purpose of this project is to store on ENSMA servers all the information on planes flying over the school.

The second data represent data stream from neOCampus project supported by the University of Toulouse III [24]. Three goals are identified for this project: ease the life of campus users, reduce the ecological print, control the energy consumption. The campus is seen as a smart city where several thousands of data streams come from heterogeneous indoor and outdoor sensors (CO2, wind, humidity, luminosity, human presence, energy and fluids consumption...).

In this section, we will illustrate the implementation result of the studied algorithms; the experiments have been done on the same set of data.

5.1. Result of quantifiers approach

In our case, we have chosen four relative quantifiers to test the database: Most, Some, Around half and Few. The definition of "Most" are given in equation (3), the relative quantifier “Some” can be defined as

$$\mu_Q(x) = \begin{cases} 10x - 1 & 0.1 \le x \le 0.2 \\ 1 & 0.2 \le x < 0.3 \\ -5x + 2.5 & 0.3 \le x < 0.5 \\ 0 & else \end{cases} \quad (5)$$

we also defined the quantifier “Around half” as following:

$$\mu_Q(x) = \begin{cases} 5x - 1.5 & 0.3 \le x \le 0.5 \\ 1 & x = 0.5 \\ -5x + 3.5 & 0.5 \le x < 0.7 \\ 0 & else \end{cases} \quad (6)$$

we can define the last quantifier “Few”, as

$$\mu_Q(x) = \begin{cases} 1 & 0 \le x \le 0.1 \\ -10x + 2 & 0.1 \le x < 0.2 \\ 0 & x \ge 0.2 \end{cases} \quad (7)$$

Figure 2 represents the membership functions of the relative quantifiers (few, some, around half and most)

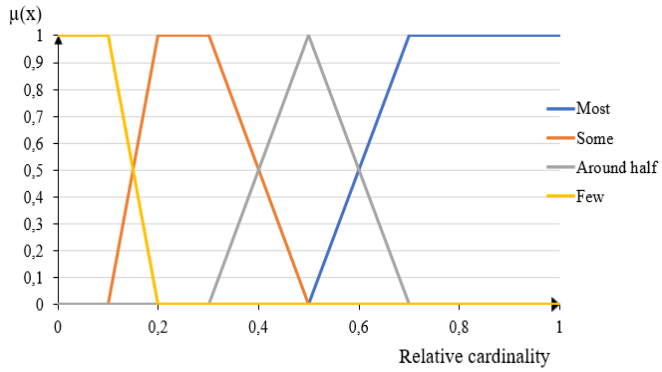

Figure 2. The relative quantifiers

A. *Static Database*

We apply quantifiers test on the flight data from ADSB project of ENSMA stored in PostgreSQL, we have chosen two attributes, the altitude of the aircraft (ALT) measured in ft, the second attribute represents the Ground Speed of the aircraft measured in knots(GS). For the first variable, we use 4 linguistic values (low, medium, high and very high) as showing in figure 3. figure 4 describes the linguistic values used for ground speed. The chosen attributes are considered among the most critical attributes during a flight, they are dependent on several other attributes such as pressure, wind speed, angle of attack.

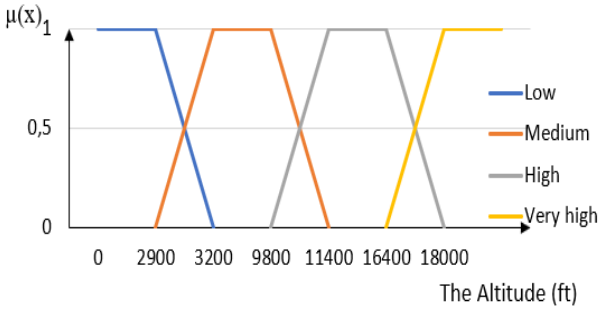

Figure 3. Fuzzy sets representing the Altitude

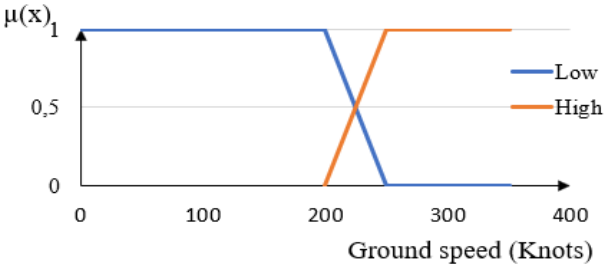

Figure 4. Fuzzy sets representing the Ground speed

The database used to conduct the experiments contains one million tuples. We get the following results:

- T (Few of Altitude are very high) = 1 which represents the best summary for the attribute Altitude.
- T (Few of ground speed are low) = 0.87 which represents the best summary for the attribute Ground speed.
- We apply quantifiers test on the flight recorder data which takes about 3.328s as execution time.

TABLE III. ALTITUDE SUMMARY EVALUATION

Quantifiers / Summarizer	Most	Around half	Some	Few
Low	0	0	0	1
Medium	0	0	0	1
High	1	0	0	0
Very high	0	0	0.91	0.09

B. *Data stream*

The data collected from neOCampus project are considered as Data stream, this later is a sequence of structured data that can be considered as infinite elements generated continuously at a fast and sometimes variable rate [12]. Since a data stream is infinite, it is not materially possible to apply processing to it as a whole. Therefore, it is necessary to define a portion of the stream to which treatment will relate and which is called a window. There are different ways of defining a window over a data stream, if we use physical time (a window of 05/29/05 at the current time for example) we speak about physical or temporal window. In our experimentation, we use the sliding window. For example, a time-based sliding window with length = 10 min produces window instances that cover the data in the last 10 minutes.

We apply quantifiers test on the temperatures collected from the smart city of the project neOCampus measured in (°c) where the distribution of the linguistic labels used to describe the temperature is represented in figure 5, the variation of these temperatures for half an hour is shown in figure 6.

- T (Most of temperature are medium) = 0.84
- T (Few of temperature are low) = 1 represents the best summary for the attribute Temperature

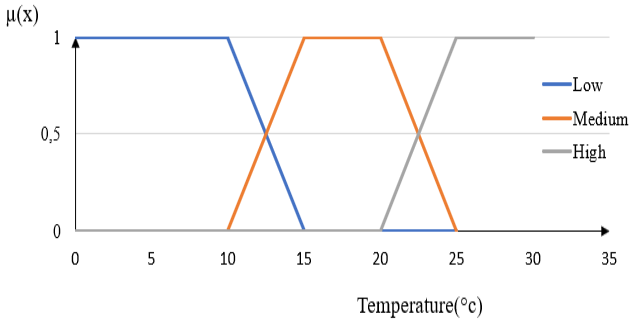

Figure 5. Fuzzy sets representing the

The use of two quantifiers “Most’’ and “Few” allows us to draw the following properties:

- **Non-contradiction:** it means that if a summary has a great degree of truth, the opposite of this summary must have a low degree of truth, taking as an example the case of the altitudes the first sentence "most of altitudes are high" has a degree which equals 1. Their negation "few of altitudes are high" has 0 as truth degree. One can also notice that the sentences have a complementary truth degree.
- **Double negation:** a summary that possesses the negation of two parameters of another summary must give us the same degree of truth as the first for example for both summary "most of the altitude are high" and "few of altitude are low" the degree of truth is 1.

Results of the typical value

In order to get the typical value from the studied data, we provide in this paper a part of the results after developing and executing algorithm 1.

A. *static database*

For the purpose of executing the algorithm, the data must be stored in a single table. The first run provides the typical value of the set of altitude (ALT) as an interval [27000, 41025].

B. *Data stream*

We apply the algorithm 1 to determinate the typical values of the temperature for different windows, the result obtained confirms that the typical value of the temperature is [21,23].

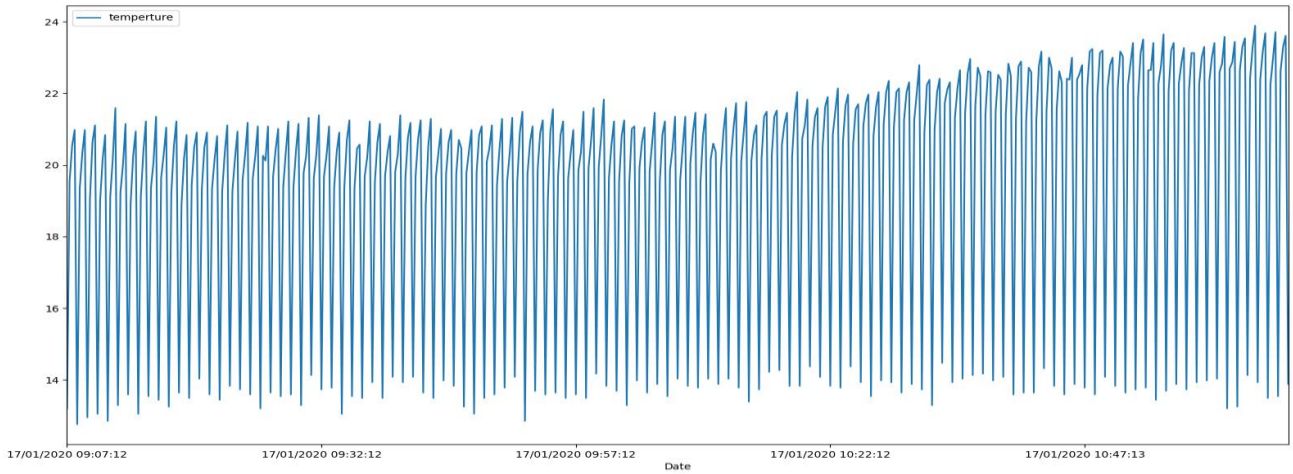

Figure 6. Temperature variation

According to the experiment tests we can say that the linguistic approach has the less execution time against Dubois and Prade[9] algorithm, this later has a good performance while provides a result to reduce a big volume of data, but it has some disadvantage which are:

- the consumption of a large space of memory;
- the execution time is great, and it is increased incrementally when the size of the database is increased.

5.2. Comparative study

Two procedures for data summarization have been described, studied and applied in several contexts of data. The experiments have been done on the same dataset. In this part, we give a comparison between two summarization algorithms seen in previous sections and we discuss the results in terms of execution time.

For the static database represented by the ADSB project, the experiments show that the quantifier approach algorithm has less execution time against Dubois and Prade algorithm. Figure 7 describes the result of the execution time as a function of the database size. As mentioned in the previous section, the algorithm of typical value has the highest execution time because it has a big combination of computation.

The results of execution time for data stream represented by the attribute temperature from neOcampus project are shown in figure 8, these results are for different sizes of sliding windows. We note that the execution time of the typical value does not differ much compared to the linguistic summary, that can be justified by measuring the amount of data collected during the window which is low for a simple reason that the sensors send 4 values per minute, so we only have 240 values for an hour.

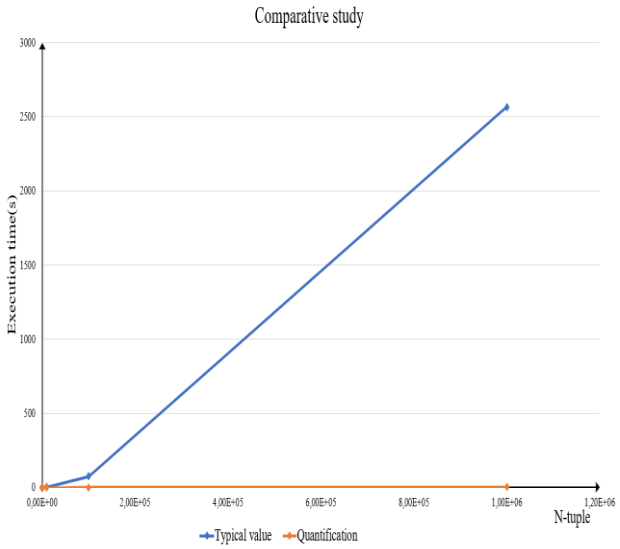

Figure 7. Comparison between the algorithms in the execution time for the static database

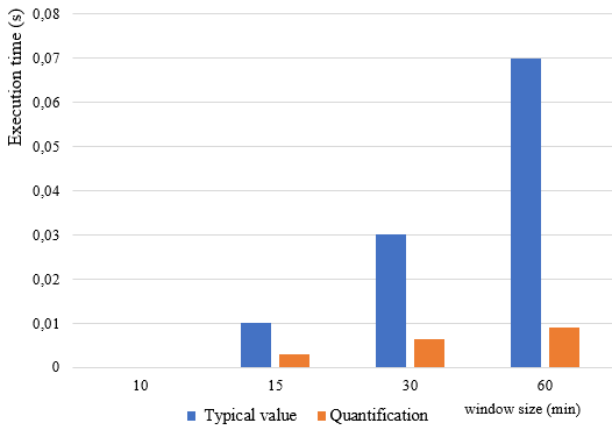

Figure 8. Comparison between the algorithms in the execution time for data stream

6. CONCLUSION

Data summarization is one of the powerful tools in knowledge extraction for large data sets. It has been used successfully in many fields. It provides the most important information in an efficient and human consistent way.

This paper has considered the summary using fuzzy logic in order to cope with the increasing volume of data created and stored. We began by studying two methods of summarization. We discussed Yager's approach based on the linguistic quantifier. The typical value has also been discussed in this paper. In the light of this study, the proposed techniques are used and implemented, a comparative study between these two concepts to obtain an efficient summary from a large set of input data was also described.

One of the challenges that will be addressed as future work is the handling of huge number of produced sentences in the quantifiers approach. Another direction for future work is to propose a typology of relevant queries on both linguistic quantifiers and typical values techniques. The evaluation of the different queries is also an interesting aspect of this work.

REFERENCES

[1] M. Ahmed, "Data summarization: a survey," Knowledge and Information Systems, pp. 1-25, 2019.

[2] F. E. Boran, D. Akay, and R. R. Yager, "An overview of methods for linguistic summarization with fuzzy sets," Expert Systems with Applications, vol. 61, pp. 356-377, 2016.

[3] M. Hudec, "Fuzziness in information systems," Springer International Publishing, 2016.

[4] R. R. Yager, "A new approach to the summarization of data," Information Sciences, vol. 28, pp. 69-86, 1982.

[5] R. R. Yager, "On linguistic summaries of data," Knowledge discovery in databases, 1991.

[6] J. Kacprzyk, "Intelligent data analysis via linguistic data summaries: A fuzzy logic approach," in Classification and Information Processing at the Turn of the Millennium, ed: Springer, 2000, pp. 153-161.

[7] J. Kacprzyk and S. Zadrozny, "Data mining via linguistic summaries of data: an interactive approach," in Methodologies for the Conception, Design and Application of Soft Computing. Proc. of IIZUKA, 1998, pp. 668-671.

[8] J. Kacprzyk and S. Zadrożny, "Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools," Information Sciences, vol. 173, pp. 281-304, 2005.

[9] D. Dubois, H. Prade, and E. Rannou, "An improved method for finding typical values," in IPMU: information processing and management of uncertainty in knowledge-based systems (Paris, 6-10 July 1998), 1998, pp. 1830-1837.

[10] B. Csernel, "Résumé généraliste de flux de données," Paris, ENST, 2008.

[11] C. D. A. Midas, "Résumé généraliste de flux de données," 2010.

[12] R. Chiky, "Résumé de flux de données ditribués," 2009.

[13] N. Gabsi, "Extension et interrogation de résumés de flux de données," 2011.

[14] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," Journal of Computer and system sciences, vol. 58, pp. 137-147, 1999.

[15] O. H. Hamid, "A model-based markovian context-dependent reinforcement learning approach for neurobiologically plausible transfer of experience," International Journal of Hybrid Intelligent Systems, vol. 12, pp. 119-129, 2015.

[16] J. Kacprzyk and S. Zadrozny, "Protoforms of linguistic database summaries as a human consistent tool for using natural language in data mining," in Software and Intelligent Sciences: New Transdisciplinary Findings, ed: IGI Global, 2012, pp. 157-168.

[17] L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages," in Computational linguistics, ed: Elsevier, 1983, pp. 149-184.

[18] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making," IEEE Transactions on systems, Man, and Cybernetics, vol. 18, pp. 183-190, 1988.

[19] J. Kacprzyk and R. R. Yager, "Linguistic summaries of data using fuzzy logic," International Journal of General System, vol. 30, pp. 133-154, 2001.

[20] D. Dubois and H. Prade, "On data summarization with fuzzy sets," in Fifth IFSA Congress, 1993, p. 465-468.

[21] R. R. Yager, "A note on a fuzzy measure of typicality," International journal of intelligent systems, vol. 12, pp. 233-249, 1997.

[22] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI magazine, vol. 17, pp. 37-37, 1996.

[23] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," in KDD, 1996, pp. 82-88.

[24] https://www.irit.fr/neocampus/fr/

Khedidja BOULANOUAR was born in Saida, Algeria, on 05 August 1992. She received the Master degree in Aeronautical Engineering from the Aeronautics Institute of Blida, Algeria, in 2016. In this same year, she started her doctoral studies in the field of aircraft sensor data management in the Aeronautics and spatial studies Institute of SAAD DAHLAB Blida 1 University - Blida, Algeria. In 2018, she started her PHD studies in avionics informatics in the national school of aeronautics ENSMA, Poitiers, French. Her current research activities include Data Reduction, Big Data Processing, Soft Computing.

Allel HADJALI is Full Professor in Computer Science at the National Engineering School for Mechanics and Aerotechnics (ISAE-ENSMA), Poitiers, France. He is a member of the Data & Model Engineering research team of the Laboratory of Computer Science and Automatic Control for Systems (LIAS/ISEA-ENSMA). His main area of research falls within Data Science field, and more specifically, the research topics related to Data Exploitation & Analysis, Knowledge Extraction and Recommendation. His current research interests include Soft Computing and Computational Intelligence in Databases, Data Uncertainty, Cooperative/Intelligent Databases, Recommendation Systems, Web Services, Approximate/Uncertain Reasoning with applications to Artificial Intelligence and Information Systems. His recent works were published in well-known journals (e.g., Applied Soft Computing, Knowledge and Information Systems,

Fuzzy Sets and Systems, International Journal of Intelligent Systems, Journal of Intelligent Information Systems and Annals of Mathematics and Artificial Intelligence). He also published several papers in International Conferences (ESWC, ICTAI, Fuzz-IEEE, FDEXA, QAS, SUM, ISMIS, IPMU, CoopIS, IFSA, ACM SAC, ICWS, SCC, ER, VLDB and EDBT (demo papers)). He co-organized several special sessions on "*Advances in Soft Computing Applied to Databases and Information Systems*" in conjunction with EUSFLAT (2009 and 2011) Conference, *"Advances in Bipolarity in Databases"* in conjunction with EUSFLAT (2013), *"Advances in Data Management in the Context of Incomplete Databases"*in conjunction with IFSA (2015) Conference, "Uncertainty in Cloud Computing" in conjunction with DEXA (2017) . He co-organized also several special issues in well-known journals, among them, *"Flexible Queries in Information Systems"* in Journal of Intelligent Information Systems (2009), *"On Advances in Soft Computing Applied to Databases and Information Systems"* in Fuzzy Sets and Systems Journal (2011), "*Post LFA 2015 Conference*" in Fuzzy Sets and Systems Journal (2017), *"Uncertain Cloud"* in International Journal of Approximate Reasoning (Starting in 2018). The complete list of his publications is available in

http://www.lias-lab.fr/members/allelhadjali.

Mohand LAGHA was born in Tizi-ouzou, Algeria, on 30 June 1976. He received the Engineer Diploma in Aeronautical Engineering from the Aeronautics Institute of Blida, Algeria, in 2000, the M.Sc. in Aeronautics Sciences from the SAAD DAHLAB University of Blida, Algeria, in 2003. He received the Ph.D. degree (with honors) in Aeronautical Engineering at the Aeronautics Department of SAAD DAHLAB Blida University on 3rd July 2008, and the habilitation (HDR) on September 2010. At present he is Full Professor in Aeronautics and spatial studies Institute of SAAD DAHLAB Blida 1 University - Blida, Algeria. His current research activities include estimation theory, radar signal processing, weather radar signal analysis, UAV applications and Big Data Processing.