Business Intelligence Reporting by Linguistic Summaries for Smart Cities: A Case on Explaining Bicycle Sharing Patterns

Erika Mináriková[1][a], Galena Pisoni[2][b], Bálint Molnár[3][c] and Hanna Kristín Skaftadottir[4][d]
[1]*Faculty of Economic Informatics, University of Economics in Bratislava, Bratislava, Slovakia*
[2]*York Business School, York St. John University, Lord Mayor's Walk, YO31 7EX, York, U.K.*
[3]*Eötvös Loránd University, ELTE, IK Pázmány Péter 1/C, 1117, Budapest, Hungary*
[4]*Department of Business, Bifrost University, Iceland*

Keywords: Linguistic Summaries, Smart Cities, Enterprise Architectures, Business Intelligence Reporting, Drill-Down and Roll-up Summaries.

Abstract: An increasing number of intelligent urban services rely on the use of Information and Communication Technologies (ICT). Data-driven approach is often considered for supporting sustainable cities, provided the pervasive nature of the Internet of Things (IoT) like sensors, and their capabilities to collect data for elaborating to the cities. This paper focuses on an intelligent business reporting approach explaining the bicycle sharing patterns by linguistic summaries in order to provide relevant insights for decision makers and citizens. We explored the developments in bicycle sharing stations in different periods of the day for months and seasons. The business intelligence query operations of drill-down and roll-up are often used in data reporting and analysis. In this work, these operations are realized by linguistic summaries. The main aim is to propose an approach for analysis and visualization in an understandable and interpretable way for diverse user categories. Experiments were conducted on the Dublin bicycle sharing data set. Finally, a way how cities can set in place the collection of data coming from different sources, as well as relevant enterprise infrastructures and data analytic pipelines for such service are discussed.

1 INTRODUCTION

The fast and undeniable growth of urban mobility services put us in the front of new challenges. New sustainable and on-demand mobility solutions have unpreceded growth, producing therefore also lots of data to work within this field (Kong et al., 2020; Mayer-Schönberger and Cukier, 2013; Provost and Fawcett, 2013). These developments require: on one hand (i) adequate enterprise architectures and solutions to properly transform and processes data coming from multiple sources, and therefore provide the basis for data analysis, and on the other hand (ii) adequate solutions for data analysis and interpretations of data to diverse respondents groups.

Smart cities should not only offer sustainable mobility services, but also work towards their successful implementation and optimization of resources allocated to them (Lim et al., 2018; Torre-Bastida et al., 2018). In this work, we tackle the problem of bicycle sharing, and the explaining use of bicycle stations that recently appear in different cities as a possible solution for the mobility problems (Midgley, 2009; Midgley, 2011). Various contemporary questions / problems that cities should tackle for the mobility services exists. For instance, which stations are usually very busy? In which time of the day considered bicycle stations are full? Does the number of tourists influence the availability of bicycles? Are bicycle stations less busy during the summer? The answers to such questions support the local authority in improving services by, for instance, adjusting capacities of the bicycle stations. For this task, the relevant data should be collected, analyzed and presented in an understandable way for local authorities (to manage improvements - adjust the capacity of the stations accordingly), citizens (to understand changes), journalists (to report developments), etc.

Digitization and data processing developments of

[a] https://orcid.org/0000-0002-4230-2109
[b] https://orcid.org/0000-0002-3266-1773
[c] https://orcid.org/0000-0001-5015-8883
[d] https://orcid.org/0000-0001-5228-8294

Mináriková, E., Pisoni, G., Molnár, B. and Skaftadottir, H.
Business Intelligence Reporting by Linguistic Summaries for Smart Cities: A Case on Explaining Bicycle Sharing Patterns.
DOI: 10.5220/0012748200003690

In *Proceedings of the 26th International Conference on Enterprise Information Systems (ICEIS 2024) - Volume 2*, pages 762-768
ISBN: 978-989-758-692-7; ISSN: 2184-4992

the past decades have transformed the field of smart cities and urban mobility profoundly. Diverse data sources are available and therefore can be handled for improving informativeness and support decision making. Internet of Things (IoT)/sensors, and information systems of the transport enterprises are examples of the mobility data.

One of the aspects of human reasoning and decision is searching for information that is not immediately seen in the collected data (Trillas, 2015), preferably in an understandable way. When we focus on the interpretation of information from data, we should bear in mind that diverse urban stakeholder categories differ in the levels of the statistical and IT literacy (Hudec et al., 2018). Thus, the information should be digested and interpreted in the most suitable way, e.g., by linguistic summaries.

In this contribution, we explore the concept known as linguistic summaries (see, e.g., (Boran et al., 2016)) to reveal, whether these summaries are beneficial and how we can improve business intelligence reporting with the linguistic summaries supporting the usual business intelligence operations of *drill-down* and *roll-up*.

From what discussed above, we set the following research questions

- RQ1. How can relevant enterprise architectures be set to collect and transform urban mobility data coming from different sources?
- RQ2. Could short quantified sentences improve explainability of data?

The article is organized as follows. Section 2 briefly explains preliminaries of linguistic summaries and data set used for experiments. Section 3 is devoted to the procedure, experiments and mining summarized sentences from the data set, whereas Section 4 is devoted to the enterprise architecture for supporting summaries. Section 5 discusses obtained results, perspectives and future tasks. Finally, Section 6 concludes the article.

2 METHODS, METHODOLOGY AND DATA SET

This section introduces linguistic summaries and data set.

2.1 Preliminaries of Linguistics Summaries

Linguistic summarization of data is a topic which occupies scientists and practitioners since Yager's seminal work (Yager, 1982). Linguistic summaries have been improved and applied in diverse fields, e.g., (Boran et al., 2016; Smits et al., 2018; van der Heide and Trivino, 2009; Wilbik et al., 2020). A summary like: *in the morning the bike station is very busy*, or *the most of young citizens commute large distances to offices* is understood at first glance. Linguistic summaries have not been applied only to interpret data, but also for revealing dependencies between data and satellite images in smart cities images (Hudec et al., 2020) among others.

Two main structures of the classic prototype forms are so-called basic structure of linguistic summaries (LS) and structure with restriction (Lesot et al., 2016). The basic structure is *Q records have S*. Quantifier Q and summarizer S are usually formalized by fuzzy sets.

The proportion (relative cardinality) of entities in a data set X that fully and partially satisfies the summarizer (predicate) S is (Yager, 1982)

$$y_{LS}(X) = \frac{1}{n}\sum_{i=1}^{n}\mu_S(x_i) \tag{1}$$

where n is the number of entities and the membership function μ formalizes summarizer S. The validity (truth value) of the summary is calculated as

$$v_{LS}(X) = \mu_Q(y_{LS}(X)) \tag{2}$$

where the function μ_Q formalizes fuzzy relative quantifier Q for the summary.

The structure with restriction is *Q R records have S*. The proportion of entities in data set that meet the summarizer S and restriction R is (Yager, 1982)

$$y_{LS}(X) = \frac{\sum_{i=1}^{n} t(\mu_S(x_i), \mu_R(x_i))}{\mu_R(x_i)} \tag{3}$$

where the membership functions μ formalize summarizer S and restriction R. The validity (truth value) of the summary is calculated as (2).

2.2 Datsets

For our experiments we used the open data set of the bicycle sharing in Dublin, Ireland. This data set is accessible at https://data.smartdublin.ie/dataset/dublinbikes-api.

The considered data set contains collected data related to the accessibility of bicycles on the bicycles station points within the city of Dublin on five minute interval for the year 2021. The following attributes are available:
station id, date time, last updated, name, bike stands, available bike stands, available bikes, status, address, latitude, longitude.

In the pre-processing steps the time was extracted from the *date time* attribute and adjusted for the linguistic terms explained later on. The availability of bicycles was calculated as a ratio between the bicycle stands capacity and and the available bicycles for every collected record. The Python libraries *numpy pandas* and *datetime* are used for prepossessing. All calculations of summaries are also realized in Python.

3 SUMMARISING BICYCLE SHARING DATA

This section focuses on summarizing data and reporting revealed patterns compatible to the usual business intelligence queries.

3.1 Procedure

Generally, LSs provide validity of any summary posed on data. However, the usual business intelligence queries supports operations of *roll up* and *drill down* (Kimball, 2011). The former gives a global overview, e.g., for top managers or shareholders, whereas the latter provides details on lower hierarchical level, e.g., for a region with a poor behavior to see, which districts are the most problematic ones, or whether poor behavior is in all districts.

In our procedure for summarising bicycle sharing data, we focuses on summarizing developments in stations by seasons. The *roll up* operation gives a global overview of a station for citizens or journalists for the entire year. The *drill down* operation is applied for seasons without a recognized pattern. The aim is to reveal, whether we can find patterns on the months or days levels.

3.2 Experiments

In order to proceed with revealing all relevant linguistic summaries from the afore explained data set (Section 2.2), we defined fuzzy sets for linguistic terms *morning*, *around lunch*, *evening* and *night* on the time attribute as is shown in Figure 1. We emphasize, that these fuzzy sets cover an usual vague separation of these parts of the day. When in a particular city or region is a different meaning of these terms, fuzzy sets can be adjusted accordingly.

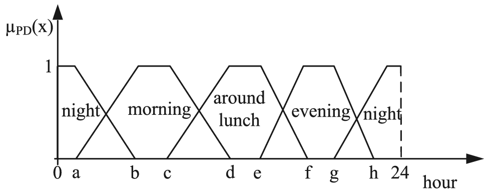

Figure 1: Part of days fuzzified into four fuzzy sets.

The next linguistic variable is availability of bicycle in stations. The number of bicycle stands in stations differ. Thus, instead of of number of bicycles we adopted the proportion of available bicycles, which is the ratio of the number of stands and available bicycles. This ratio is fuzzified into three fuzzy sets *few*, *about half* and *most of* bicycles are available. The next required concept is an elastic quantifier *most of* to express sentence like *the most of mornings (in the considered time period) few bikes is available*.

In the next step, we executed summaries for each combination of time and bicycle availability and select the most relevant quantified sentences. We calculated summaries for two stations. For the other stations, the calculations are straightforward. In order to shorten sentences, we excluded quantifier *most of*, i.e., instead of *the most of mornings few bikes is available* we write *mornings is available few bikes*. The results (the sentences with the highest validity) for stations ID(117) and ID(38) are as follows.

Season:01-04 (Winter)
Station ID(117)
morning is available few_bikes t= 0.9748
around_lunch is available few_bikes t= 0.9875
evening is available few_bikes t= 0.9852
night is available few_bikes t= 0.9820

Season:01-04 (Winter)
Station ID(38)
morning is available around_half_bikes t= 0.5023
around_lunch is available few_bikes t= 0.5277
evening is available around_half_bikes t= 0.5635
night is available around_half_bikes t= 0.5372

Season:07-10 (Summer)
Station ID(117)
morning is available few_bikes t= 0.9592
around_lunch is available few_bikes t= 0.9740
evening is available few_bikes t= 0.9666
night is available few_bikes t= 0.9628

Season:07-10 (Summer)
Station ID(38)
morning is available many_bikes t= 0.3975
around_lunch is available few_bikes t= 0.4227
evening is available around_half_bikes t= 0.4764
night is available many_bikes t= 0.4494

Summaries revealed that the station ID(117) is very busy during both seasons, whereas for station ID(38) the dominant bicycles availability pattern does not exist. Thus, we try by the *drill down* operation to the months level to find, whether we can recognize patterns in respective months.

Drill down to months (Station ID 38 summer):
July
morning is available many_bikes t= 0.4246
around_lunch is available around_half_bikes t= 0.451
evening is available many_bikes t= 0.5129
night is available many_bikes t= 0.5189

No recognised dominant pattern, we continue with drill down on days:

01.07.2021
morning is available around_half_bikes t= 0.0
morning is available few_bikes t= 1.0

02.07.2021
morning is available around_half_bikes t= 0.8537
morning is available few_bikes t= 0.0269

03.07.2021
morning is available many_bikes t= 0.9944
morning is available around_half_bikes t= 0.0083
morning is available few_bikes t= 0.0009
etc.

This drill down operation is realizable due to the data availability for each minute.
For each day summaries recognised different behaviour. This is the reason, why with the aggregated data no dominant pattern is recognised.

August
morning is available many_bikes t= 0.4351
around_lunch is available few_bikes t= 0.4834
night is available many_bikes t= 0.4671

September
morning is available few_bikes t= 0.3718
around_lunch is available few_bikes t= 0.4597
evening is available around_half_bikes t= 0.5172

No significant pattern for both months - another drill down is required.

October
morning is available many_bikes t= 0.5701
around_lunch is available around_half_bikes t= 0.5968
evening is available many_bikes t= 0.9583
night is available many_bikes t= 1.0

In the last month, we recognized that many bicycles are available in evening and night.

The opposite operation in roll up to the year 2021 for this station (ID38):
morning is available around_half_bikes t= 0.382
around_lunch is available around_half_bikes t= 0.4699
evening is available around_half_bikes t= 0.4722
night is available around_half_bikes t= 0.4175

Again, not a dominating pattern is recognized. In our experiments, we considered pattern to be dominant when its validity is greater than or equal to 0.75.

For station ID117 there is not a significant difference between the considered sessions. The revealed patterns indicate that constantly only few bicycles is available. This station is very busy during the entire year. Thus, the increasing in capacity should be considered, or at least the focus of business intelligence dashboard should be on this station.

Contrary, station ID38 is rarely busy. Thus, stands can be reduced and moved to another station. The most relevant sentence has validity slightly above 0.5.

4 ENTERPRISE ARCHITECTURE FOR DATA COLLECTION AND MINING SUMMARIES

Recently, the fields of smart mobility became one of the primary sources of data through the applications of various sensors and IoT-s. In the ecosystem of smart cities and its smart mobility, the efficient and effective utilization of data become an essential issue. Transformation of data collection from the simple substantiated IoT data to data that originate from large-scale monitoring led to the requirement of disciplined data analytic. Data can be listed as a traditional data coming from the city operation, and operational data of transport originated from different devices. The latter data are usually unstructured and heterogeneous; either we consider their structure or their content. Data should be accompanied by metadata that describe the essential information about the content and can be utilized to categorize and organize data to exploit them for the advanced data mining.

When focusing on the bicycle sharing, it depends on various aspects like weather, restrictions (e.g., pandemic or events). Such data are usually not available in the data collected from sensors on the stations.

Previous research has shown that analysis of data that originated from structured databases and external sources of a clear relation to structured data is relatively straightforward with the use of data warehouses (Golfarelli and Rizzi, 2009; Kimball, 2011; Kimball and Ross, 2013). In the bicycle sharing, such data can be from the meteorological data sets explaining weather conditions near each station, or whether events appear near the stations. The essential function of the data warehouse is to integrate data and transform them into the confirmed structure, relevant for the domain of interest. In this case, the domain of the urban sustainable mobility to support reporting and interpreting for the diverse urban stakeholders groups.

In Section 3.2, linguistic summaries are applied on the bicycle sharing data to reveal patterns. Beneficially, the Key Performance Indicator (KPI) for the stakeholders should be expressed linguistically (Vaisman and Zimányi, 2022). The reporting with linguistic summaries gives a reliable, trustable, and easy-to-understand overview of the actual state regarding the stations (see Section 3.2). This interpretations provide the basis for supporting effective decision-making for the city (see Figure 2). In our case, KPI is expressed by linguistic terms: *few*, *about half* and *most of*.

In order to generate the linguistic insights as afore computed, the relevant ”data ingestion” framework, or the way how the data are fed to the data warehouse, should be set in place. Relevant levels of security should be treated in line with existing regulation (foremost GDPR) for data management.

The benefit of implementing an approach as proposed in this work, is that a local authority can, if of interest, extend the reporting of the collected data by the other internal and external data and adjust reporting to diverse user categories. For instance, citizens and users with disabilities benefit of LS, which provide an easy way to communicate mined patterns. For advanced users, like traffic experts, external data covering, e.g., traffic densities of cars, weather conditions and organized events can reveal how these parameters influence bicycle sharing in the affected stations. With an approach as is under the development in this work, local authorities would have a better overview of cycling in the city. Next, suited LSs can be adopted for informing cyclists about their riding behaviour in comparison to the average values, for instance. However, a summary of structure *your length of ride is around average* is not informative enough. This summary is the same for person who use bicycle every working day (but not during weekends) and for someone who rarely cycle on working days, but is a heavy bicycle user during weekends. The better option are quantified linguistic summaries like *few days your length of rides is slightly under average, and about half of days your length of rides is about average.*

5 DISCUSSION

The proposed approach has a significant applicability potential. This holds especially when informing diverse urban stakeholders groups (including disabled citizens) should be realized by a robust and compact approach.

In our data intense society, we face the problems of explaining mined patterns (Smits et al., 2018). A promising way is by linguistic summaries. In order to contribute, we raised two research questions. The first research question is *How can relevant enterprise architectures be set to collect and transform urban data coming from different sources*? The answer is that the processing data by the linguistics summaries requires a clear data warehouse model for storing data coming from diverse sources. Such data are relevant for advanced explaining summaries and revealed patterns.

The second research question is *Could short quantified sentences improve explainability of data*? The answer depends on the structure of the sentences. If it is a basic structure of summary (1), then histograms and the other charts are suitable when visual attention is not disturbed (i.e., it should not be focused elsewhere) (Arguelles and Triviño., 2013), or summary is not for visually impaired citizens. Regarding, the summary with restriction, the situation is different. It is more convenient to express it by sentences like in Section 3.2 than on series of graphs covering more attributes. Anyway, linguistically summarized sentence is convenient for all urban stakeholders groups.

The next perspective is merging summaries with maps. Either by using map features in summarized sentences (Hudec et al., 2020), like whether a particular proportion of objects influences the other attributes, or interpreting summarized sentences on maps. The positions of two stations used in experiments (Section 3.2) are shown on map in Figure 3. The very busy station is located near the river and canal.

The main problem of mining all relevant summarized sentences is the computational cost, when we consider all stations and all possible summaries in a large city. In the future work, we focus our work on the optimisation in this direction.

6 CONCLUSIONS

A smart and sustainable mobility should rely on a heavier use of bicycles. One possibility is by the bicycle sharing concept. This work has shed light on summarizing the bicycle sharing patterns of the bicycle stations.

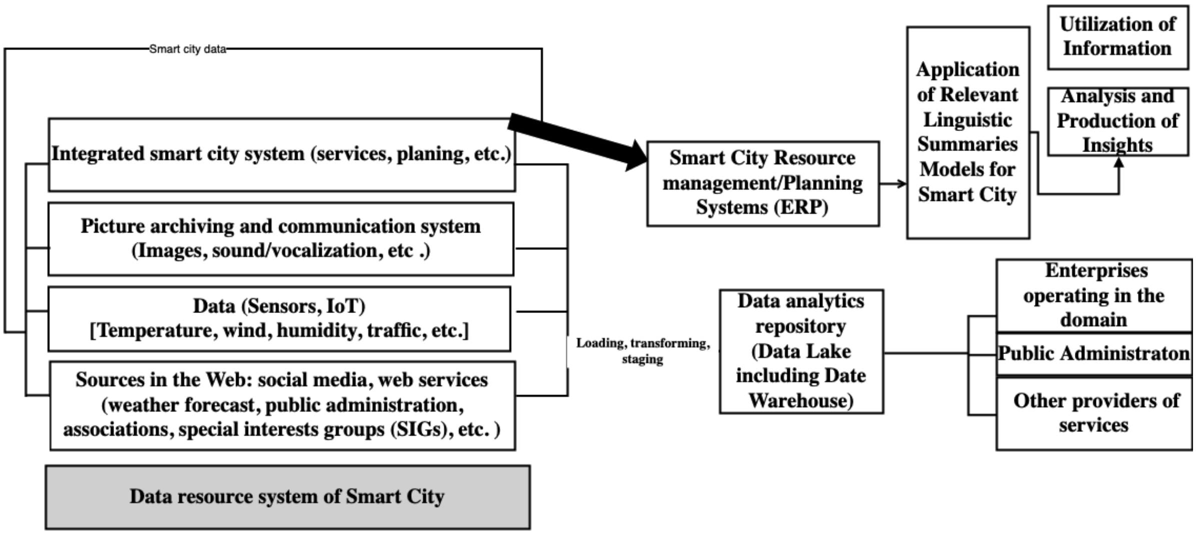
Smart city data
Integrated smart city system (services, planing, etc.)
Picture archiving and communication system (Images, sound/vocalization, etc .)
Data (Sensors, IoT) [Temperature, wind, humidity, traffic, etc.]
Sources in the Web: social media, web services (weather forecast, public administration, associations, special interests groups (SIGs), etc.)
Data resource system of Smart City
Loading, transforming, staging
Smart City Resource management/Planning Systems (ERP)
Application of Relevant Linguistic Summaries Models for Smart City
Utilization of Information
Analysis and Production of Insights
Data analytics repository (Data Lake including Date Warehouse)
Enterprises operating in the domain
Public Administraton
Other providers of services

Figure 2: Sources of data and usage, figure composed by the authors, the creation of the figure was inspired by (Kimball, 2011; Mayer-Schönberger and Cukier, 2013; Provost and Fawcett, 2013; Kong et al., 2020).

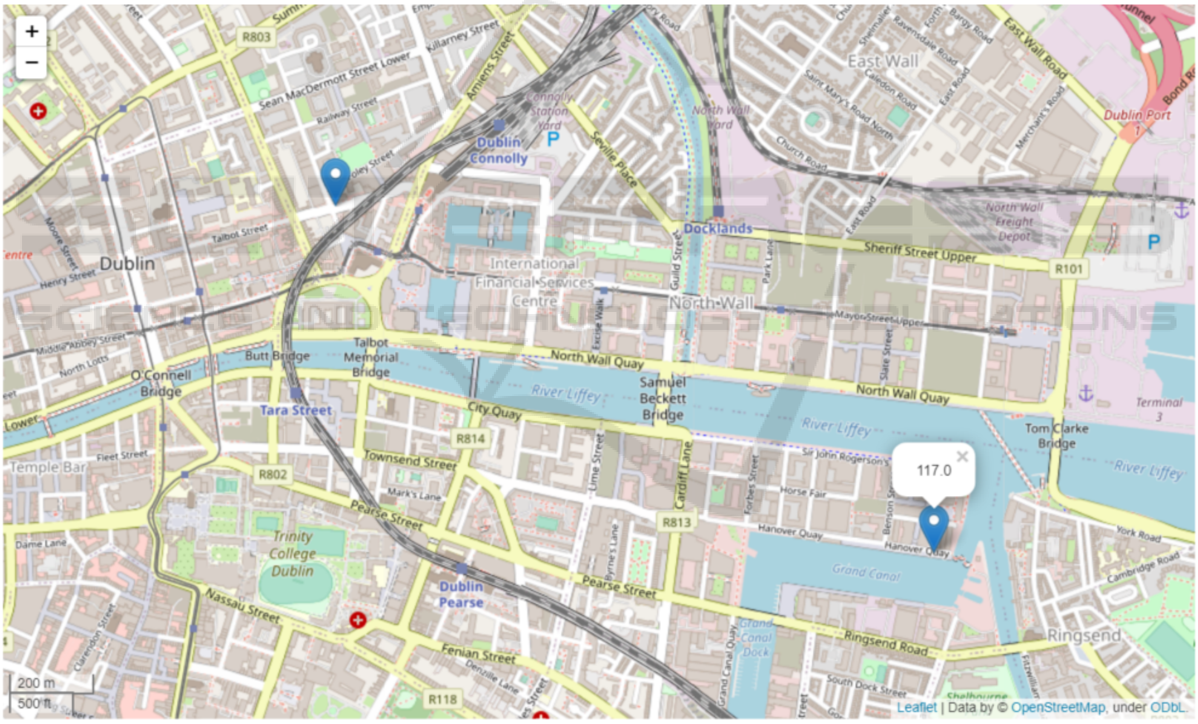

Figure 3: Location of the two stands used in experiments.

In business intelligence reporting a well-known way is by *drill down* and *roll up* query operations (e.g., in MDX query language). In this work, we created these operations by linguistic summaries. Next, we demonstrated summaries (including *drill down* and *roll up*) on the real world open dataset coming from the bicycle sharing service in Dublin, Ireland.

In this paper, we also proposed an architecture design required for collecting and transforming data coming from different sources as a support for revealing and interpreting patterns by linguistic summaries. It is an explainable way how to provide insights into the bicycle sharing data, a relevant support for decision makers, local authorities and citizens.

This position work requires further investigation in the field of effective data collection and integration due to diversity of required data, as well as in the optimization of mining relevant patterns of bicycle sharing by linguistic summaries. Linguistic summarization copes with the high computational demand due to a larger number of data expressing bicycle sharing stations and variety in summarized sentences.

ACKNOWLEDGEMENTS

This research was supported the Thematic Excellence Programme TKP2021-NVA-29 (National Challenges Subprogramme) funding scheme, the COST Action CA19130 - ”Fintech and Artificial Intelligence in Finance Towards a transparent financial industry” (FinAI), and the FBR-PDI-021 - The cooperation in the fields of business intelligence and artificial intelligence for science and education - BICISEDU.

REFERENCES

Arguelles, L. and Triviño., G. (2013). I-struve: Automatic linguistic descriptions of visual double stars. *Engineering Applications of Artificial Intelligence*, 26:2083–2092.

Boran, F., Akay, D., and Yager, R. (2016). An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications*, 61:356–377.

Golfarelli, M. and Rizzi, S. (2009). A survey on temporal data warehousing. *International Journal of Data Warehousing and Mining (IJDWM)*, 5(1):1–17.

Hudec, M., Bednárová, E., and Holzinger, A. (2018). Augmenting statistical data dissemination by short quantified sentences of natural language. *Journal of Official Statistics*, 34:981–1010.

Hudec, M., Vučetić, M., and Čermáková, I. (2020). The synergy of linguistic summaries, fuzzy functional dependencies and land coverings for augmenting informativeness in smart cities. In *28th Telecommunications forum TELFOR 2020*. IEEE.

Kimball, R. (2011). The evolving role of the enterprise data warehouse in the era of big data analytics. *Whitepaper, Kimball Group, April*.

Kimball, R. and Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling (3rd ed.)*. Wiley, New York.

Kong, L., Liu, Z., and Wu, J. (2020). A systematic review of big data-based urban sustainability research: State-of-the-science and future directions. *Journal of Cleaner Production*, 273:123142.

Lesot, M.-J., Moyse, G., and Bouchon-Meunier, B. (2016). Interpretability of fuzzy linguistic summaries. *Fuzzy Sets and Systems*, 292:307–317.

Lim, C., Kim, K.-J., and Maglio, P. P. (2018). Smart cities with big data: Reference models, challenges, and considerations. *Cities*, 82:86–99.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Midgley, P. (2009). The role of smart bike-sharing systems in urban mobility. *Journeys*, 2(1):23–31.

Midgley, P. (2011). Bicycle-sharing schemes: enhancing sustainable mobility in urban areas. *United Nations, Department of Economic and Social Affairs*, 8:1–12.

Provost, F. and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59.

Smits, G., Nerzic, P., Pivert, O., and Lesot, M. (2018). Efficient generation of reliable estimated linguistic summaries. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ–IEEE)*. IEEE.

Torre-Bastida, A. I., Del Ser, J., Laña, I., Ilardia, M., Bilbao, M. N., and Campos-Cordobés, S. (2018). Big data for transportation and mobility: recent advances, trends and challenges. *IET Intelligent Transport Systems*, 12(8):742–755.

Trillas, E. (2015). An algebraic model of reasoning to support zadeh's cww. In Kacprzyk, J. and Pedrycz, W., editors, *Handbook of Computational Intelligence*, pages 249–267. Springer, Berlin Heidelberg.

Vaisman, A. and Zimányi, E. (2022). *Data Warehouse Systems – Design and Implementation*. Springer-Verlag, Berlin Heidelberg.

van der Heide, A. and Trivino, G. (2009). Automatically generated linguistic summaries of energy consumption data. In *Ninth International Conference on Intelligent Systems Design and Applications (ISDA '09)*. IEEE.

Wilbik, A., Barreto, D., and Backus, G. (2020). On relevance of linguistic summaries – a case study from the agro–food domain. In *8th International Conference on Information Processing and Management of Uncertainty in Knowledge–Based Systems (IPMU 2020)*. Springer.

Yager, R. (1982). A new approach to the summarization of data. *Information Sciences*, 28(1):69–86.