A Multilevel Graph Approach for Road Accidents Data Interpretation

Fabio Clarizia[1(✉)], Francesco Colace[1(✉)], Marco Lombardi[1(✉)],
Francesco Pascale[1(✉)], and Domenico Santaniello[2(✉)]

[1] DIIn, University of Salerno, 84084 Fisciano, SA, Italy
{fclarizia, fcolace, malombardi, fpascale}@unisa.it
[2] DICIV, University of Salerno, 84084 Fisciano, SA, Italy
dsantaniello@unisa.it

Abstract. Nowadays, due to the massive low-cost technology and mobile devices spread, our society is increasingly projected towards data production. Often, we find ourselves surrounded by data that, however, does not always lead to the knowledge, or toward information that we need. This is liable to eclipse the desire to use this data trying to predict the future. So much has been done in literature in regard to the extraction of information and interpretation of these data. However, in this field does not seem to be present a universal methodology for solving the problem, leading us to research new approaches more customized on the available dataset. The aim of this paper is to introduce an approach for the interpretation of data from sensors located within a city using three graphical views: Context Dimension Tree, Ontologies and Bayesian Networks. Through the Ontologies and the Context Dimension Tree it is possible to analyze the scenario from a syntactic and semantic point of view, assisting the construction of the he Bayes network structure that allow to estimate the probability that some events happen. A first preliminary analysis conducted on a London borough seems to confirm the effectiveness of the proposed method.

Keywords: Big Data · Context Awareness · Smart City
Knowledge management

1 Introduction

An ever-increasing digitization of our lives involves the production of such a large and rich amount of data which can not be processed by conventional methods. Being able to develop solutions and algorithms capable to interpret and interact with this huge amount of information is a key challenge which Big Data poses in today's world [1]. However, many questions remain: How to conduct this transformation? How to properly use these data to increase the competitiveness and efficiency of services? And, how could they contribute to social development? [2].

Data management has grown along three dimensions: volume, velocity and variety [3]. According to Tole [4] the so-called "3Vs" represents key elements regarding the characteristics of Big Data systems. Volume refers to the amount of structured or

© Springer Nature Switzerland AG 2018
A. Castiglione et al. (Eds.): CSS 2018, LNCS 11161, pp. 303–316, 2018.
https://doi.org/10.1007/978-3-030-01689-0_24

unstructured data generated, which is being manipulated and analysed in order to obtain the desired results. Commonly, these data are generated by heterogeneous sources such as traditional databases, social media, sensors, logs, events, etc. Velocity refers not only to the speed of data generation, but also to the need, of this information, to be processed in real time. Variety deals with the different types of data that are generated, collected and used. These data, which belong to the most disparate codification, suggests the use of different storage and retrieval approaches.

These pervasive data are mainly the result of two independent phenomena that reached critical mass simultaneously: the advent of the Internet of Things and the increase in volume of user-generated content produced by social networks and smart mobile terminals [5].

The internet has allowed us to create a powerful information network, through which more and more services are spread: from information to communication, from banking services to the purchasing. In addition, it has given us the opportunity to connect human beings to each other, to communicate and share anything anywhere instantly. In this sense, in 1999 Kevin Ashton defined the term Internet of Things (IoT) [6] which refers to the concept of a network in which human beings and machines are connected, using common public services. According to Atzori [7] the main strength of the IoT idea is the high impact it will have on several aspects of everyday-life and behaviour of potential users. The spread of low-cost sensors makes a significant contribution to create an impressive amount of data. Moreover, thanks to their pervasiveness they appear able to influence our daily actions more and more. This leads us to pose further questions: How can we properly process this data? Is it necessary to understand the underlying processes generating the data precisely? How are the data sources linked with each other? Is it sufficient to have high level general views in order to mine useful results? How to turn these data into knowledge? Are there specific techniques and methodologies able to analyse this important amount of data? Much has been done regarding to manage these huge volumes of data (Hadoop, Spark, Storm, Google BigQuery, etc.), but what has been done for the interpretation of these data?

It's been a while since the literature no longer refers to the 3Vs but to the 5Vs [8], Value and Veracity are introduced as fundamental characteristics to analyse the problems related to the Big Data. Value is a key aspect of the data, which is defined by the added-value that the collected data can bring to the intended process, activity or predictive analysis/hypothesis; obviously this aspect is related to the capability to transform data into Knowledge Database. The veracity dimension of Big Data includes data consistency and data trustworthiness, related to a number of factors including statistical reliability, data origin, processing methods etc. It is important to ensure the reliability of the data considering that results may be generated on which important decisions are made. Assigning a veracity index to data on which the analyses are based is essential in order to have a measure of the general reliability of the system. These aforementioned issues are all key when dealing with data generated in a smart cities context.

Increasing populations and rapid large-scale urbanization creates a demand to increase the quality of life through economic development, environmental efficiency and stability. This could be performed by designing urban areas which take advantage of integrated technologies and the optimization of resources in order to improve some key objectives such as mobility, communication, economy, work, environment, administration and construction. It was 2008 when IBM, during the years of the global financial crisis, suggested a smart approach to deal with problems afflicting economic growth launching the concept of a smarter planet. Smart cities are able to use data such as traffic congestion, power consumption statistics, and public safety events, in order to upgrade the city services, through three foundational concepts: instrumented, interconnected, and intelligent supplies [9]. Instrumented is referring to sources of data from physical or virtual sensors, interconnected refers to the capacity of the integration and management of those data into an enterprise computing platform and their communication, Intelligent refers to the capacity of complex analytics, modelling, optimization, and visualization in order to make better operational decisions.

Many applications, on smart cities concept, have been proposed in literature; in particular, Zanella et al. present and discuss the technical solutions and best-practice guidelines adopted in the Padova Smart City project. Added-value services for citizens and the administration of the city have been highlighted in many areas of interest such as: Structural Health of Buildings, Waste Management, Air Quality, Noise Monitoring, Traffic Congestion, City Energy Consumption, Smart Parking, Smart Lighting [10].

Information management environments, or more generally pervasive data contexts, may be supported by context representation approaches and enhanced through adopting probabilistic approaches such as Bayesian Network (BN) [5]. BNs can offer a framework for risk and maintenance analysis through their ability to model data transparently. Some of the advantages of probabilistic approaches are the capability to model complex systems, to make predictions as well as diagnostics, to compute the probability of an event, to update the probabilities according to evidence, to represent multimodal variables and offer a user-friendly graphical and compact approach [11].

As previously mentioned, a further element of added value could be given by the introduction of methodologies capable to represent the context, in particular, the Context Dimension Tree (CDT). The CDT represents a valid tool used for applications which include the choices of places of interest [12]. In addition, CDT, or more generally context-aware approaches, leads to the rationalization of information delivered to the users and to the personalized distribution of information [13]. An undisputed and widely used method for representation of reality are ontologies. An ontology can adequately support pervasive context-aware systems [14], in addition, there is a strong connection with Bayesian Networks [22]. In particular, according to Helsper et al. is possible to build BNs through Ontologies [15], and vice versa Colace et al. propose a novel algorithm for Ontology building through the use of BNs [16].

Thinking about the mentioned context representation methodologies and the ability of the BNs, which from experimental evidences and through probabilistic approach are able to identify probable events, it is necessary to introduce techniques and methodologies able to manage the context in real time, in order to improve the quality of life in smart cities. The aim of this paper is to introduce a methodology for merging context representation techniques, which are CDT and Ontology, and probabilistic approach based on BN in order to help expert user to handle emergency conditions or provide suggestions for the liveability of the citizens. Sample Heading (Third Level). Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

2 The Proposed Approach

In a great number of cases, the problem we have to sort out is the following one: given a series of data, facts or observations, we are interested in identifying their most likely source and the reason that they have been generated, with a view to optimizing our own decisions. Although this seems quite a simple operation, making a decision in uncertain conditions is a process, which is far from being trivial. In this respect, the goal here is to identify an architecture to be used as an extremely flexible inferential/decision-making tool. Such architecture will not only enable the managing of complex problems, featuring a great variety of variables inter-linked through both logical-deterministic and probabilistic relationships, but also provide an effective graphic representation of the phenomenon at stake, formulating a problem description that will enhance the degree of comprehension and allow the identification of key variables. The innovative characteristics of the proposed architecture concern mainly the informational content that is intended to be made available to the end users with three point of view: Data management and organization, Representation of the context and Inferential engines.

2.1 Data Management and Representation

Data represents the key to build up and enable services and actions to be made: the goal is then to implement a Knowledge Base (KB) with a view to collecting, elaborating and managing information in real time. In this respect, we use a Knowledge Organization System (KOS), by which we mean well known schemes such as Taxonomies, Thesaurus and further types of vocabulary that, together with Ontologies, constitute valid tools to shape the reality of interest into concepts and relations between concepts [17]. Many benefits stem from this: using ontologies, for instance, allows to fix a series of key concepts and definitions relating to a given domain that can be shared, thus making the appropriate terminologies available (collaborative knowledge sharing); furthermore, an ontology allows a full re-usage of the knowledge that it codifies, even within other ontologies or rather for their completion (non-redundancy of information) and, being susceptible to interpretation by electronic calculators, enables the automatic treatment of knowledge with relevant significant advantages (Semantic Web).

2.2 Representation of the Context

The goal is primarily to deliver to different categories of users, in a given moment, information which is useful in a given context. In practice, the objective would be to set up an architecture characterized by a high degree of Context Awareness. Real time understanding of the context where users are, via a representation by means of graphs, enables the provision of a wide array of personalized, “tailored” services and suggestions regarding the decisions to make, that can help them in professional and private daily life, managing in the best possible way both the time and resources they have, hence meeting their needs [18]. Context Awareness should be understood as a set of technical features capable of providing added value to services in different operational segments. Context Aware Computing applications can exploit, in this specific case, such features in order to provide context-related information to users, or suggest them an appropriate selection of actions. In order to achieve a better representation of the various features, formal tools of context representation have been adopted, capable to define in details the user's needs in the context where he is acting, through an approach « where, why, when, how » . In detail, the representation of the context has been implemented by means of formal models of representation, such as the Context Dimension Tree (CDT). The CDT is a tree composed of a triple <r; N; A> where r indicates its root, N is the set of nodes of which it is made of and A is the set of arcs joining these nodes. CDT is used to be able to represent, in a graphic form, all possible contexts that you may have within an application. Nodes present within CDT are divided into two categories, namely dimension nodes and concept nodes. A dimension node, which is graphically represented by the colour black, is a node that describes a possible dimension of the application domain; a concept node, on the other hand, is depicted by the colour white and represents one of the possible values that a dimension may assume. Each node is identified through its type and a label. The children of the root node r are all dimension nodes, they are called top dimension and for each of them there may be a sub-tree. Leaf nodes, instead, must be concept nodes. A dimension node can have, as children, only concept nodes and, similarly, a concept node can have, as children, only dimension nodes. A Context Element is defined as an assignment dimension_namei = value, while a Context is specified as an "and" among different context elements: several context elements, combined with each other, give rise to a context.

2.3 Inferential Engines

The system collects data from various sources without interruption and immediately processes them, with a view to activating precise actions, depending on the users and on the events. These events, detected and analysed, will have to be translated into facts associated to specific semantic values: it is therefore necessary to use an inferential engine capable to draw conclusions by applying certain rules to reported facts, which could be imagined as a sequence of if-else statements. The approach selected to implement this inferential engine stems from Bayesian Networks: powerful conceptual, mathematic and application tools allowing the management of complex problems with a great number of variables interlinked by means of both probabilistic and deterministic relations.

3 The System Architecture

The system architecture, sketched out in Fig. 1, envisages functional blocks with three main phases. In the first phase, defined as the Collection Phase, data (referred to as "rough data"), are provided by different types of sensors. The set of data that are most significant with a view to the analysis that is meant to be carried out, is saved within a database. Then, in the Pre-Processing Phase, data are transformed in order to adapt them to the system that will have to use them. In general, data arrives from different sources and therefore show inconsistencies such as, for instance, the usage of different denominations to identify the same value of a feature. In addition, this phase envisions the cleaning of the collected data, in order to eliminate any error, and the treatment of missing data. The phase ends up with sampling and discretization of data. Finally, the Elaboration Phase aims at providing a representation and interpretation of the acquired knowledge, starting from information correctly memorized. To this end, an approach is followed which is based on the three views previously described, leading to implementing and using “decisional models”. Such models are constantly improved based on newly collected data and experiences, or previously treated cases.

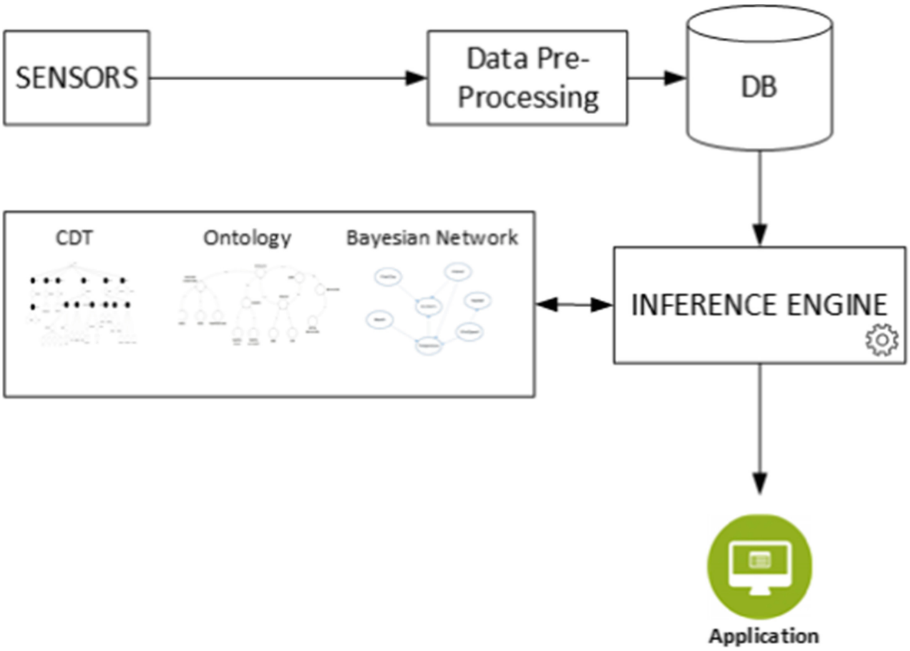


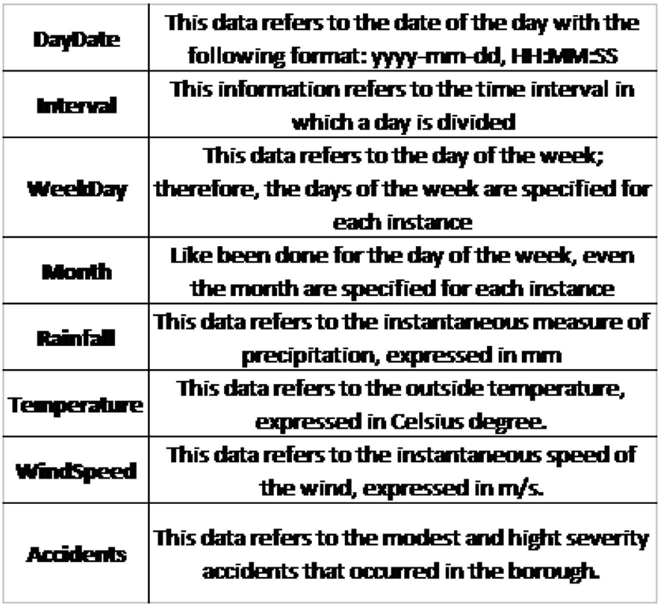
Fig. 1. The system architecture

Summing up, the need to make a decision, in a given context, can be met through the fruition of the right information delivered by the architecture. This information is featured by innovative elements based on: knowledge management and organisation, formal context representation, inferential engines.

4 Experimental Results

This section presents the experimental results of the proposed approach: the architecture is designed to collect and analyse a vast amount of data, making it available to different categories of users. The results shown are aimed at highlighting the strength of the system, which would be the ability to adapt quickly and the exploitation of human-machine interaction in order to provide automatic and reliable answers. The study area is the city of London where, for data availability reasons, it was possible to collect a sufficient number of data to provide a preliminary example that allow us to show the capability of the system to provide a reliable Bayesian Network capable to predict accident risk. In particular, the selected borough is Westminster, which is an inner London borough that occupies much of the central area of Greater London including most of the West End, and the observation period of data set is throughout the 2016 year. The data, obtained from sensors spread throughout the borough, were aggregated at 3-h intervals, resulting in the observation period of one year being made of 2920 instances. The data are organized as shown in the table below (Table 1).

Table 1. Dataset



DayDate	**This data refers to the date of the day with the following format: yyyy-mm-dd, HH:MM:SS**
Interval	**This information refers to the time interval in which a day is divided**
WeekDay	**This data refers to the day of the week; therefore, the days of the week are specified for each instance**
Month	**Like been done for the day of the week, even the month are specified for each instance**
Rainfall	**This data refers to the instantaneous measure of precipitation, expressed in mm**
Temperature	**This data refers to the outside temperature, expressed in Celsius degree.**
WindSpeed	**This data refers to the instantaneous speed of the wind, expressed in m/s.**
Accidents	**This data refers to the modest and hight severity accidents that occurred in the borough.**

To explain the capability of the presented system, the experimental phase is shown by comparing three different cases. The data set, in order to perform the analysis, is divided in a Training data Set (TrDS), which represents 75% of the data (2190 instances), and in a Test Data Set (TeDS), which represents 25% of the data (730 instances). The analysis is performed through R-Studio IDE [19]. The results of the analysis are provided in the form of Confusion Matrix, also known as error matrix, which is a specific matrix containing observed data and predicted results that allows valuation of the performance of an algorithm. Through Confusion Matrix is possible to give results in terms of Accuracy, which is a description of systematic errors, a measure of statistical bias. In this way three Bayesian Networks, obtained according to three different approaches, will be compared. The cases are the follows: Defined Bayesian Network, Learned Bayesian Network, MuG Bayesian Network.

4.1 Case #1

An expert defined BN structure, shown in Fig. 2, is taken into account, it is combined with the TrDS in order to obtain the conditional probabilities. At this point we can test the obtained BN comparing the predicted results and the observed data.

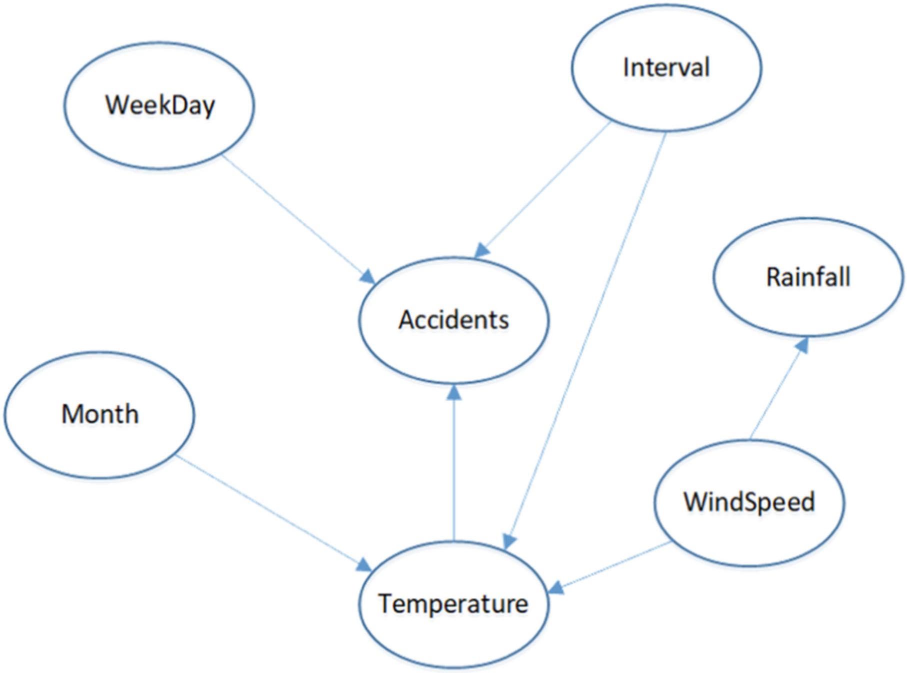


Fig. 2. Defined Bayesian network

4.2 Case #2

The BN is defined and learned with a chosen structural learning algorithm through the TrDS. The so learned network is tested with the TeDS in order to obtain the confusion matrix. The Score-based Learning Algorithm chosen is K2 Hill Climbing [20]. It has been possible to use this algorithm through the bnlearn package [21] available for the programming language R. The Bayesian Network structure is shown in Fig. 3.

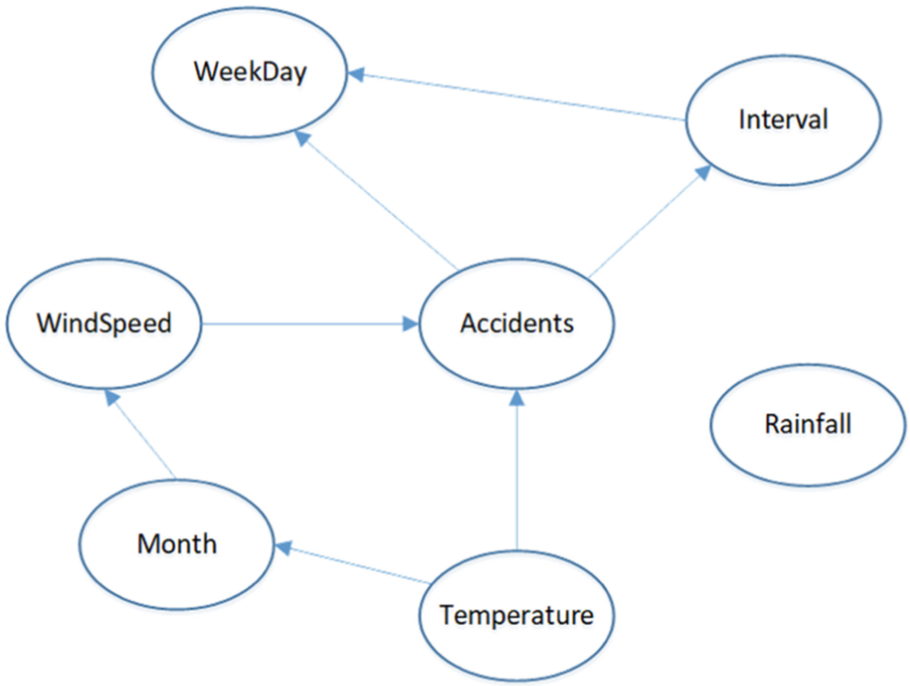


Fig. 3. Learned Bayesian network

4.3 Case#3

In this case, our approach has been applied combining CDT and Ontology in order to obtain a reliable BN. In the first phase the CDT (Fig. 5) and the Ontological view (Fig. 6) are taken into consideration. The system automatically makes a selection of all the nodes in all possible combinations according to the target. Starting by selecting some nodes of the CDT, the same nodes will be selected on the Ontological view by extracting their relationships. These relationships are turned into a constraints list, which is an essential tool in the BN building process. For example, a relation found combining CDT and Ontology is the follow:

WeatherConditions ***has_influence_on*** *Randomness*
This is automatically manipulated, obtaining:
(Rain ***is_subclass_of*** *WeatherCondition)* ***has_influence_on*** *(Randomness* ***is_subclass_of*** *RoadAccident)*
Rain ***has_influence_on*** *RoadAccident*
Finally, it is turned into a constraint:
Rainfall ***TO*** *Accidents*

The same is been done in the next example:

Event ***has*** *TemporarlThing*
(RoadAccident ***is_subclass_of*** *Event)* ***has*** *(Month* ***is_subclass_of*** *TemporarlThing)*
RoadAccident ***has*** *Month*
Finally, the constraint obtained is
Accidents ***NOT_TO*** *Month*

In the second phase, the Bayes Network is defined and learned with a preselected structural learning algorithm, which is the same of the Case #2, and the constraints list through the TrSD. In the third phase the so-learned network, which is shown in Fig. 4, is tested with TeDS in order to obtain the confusion matrix.

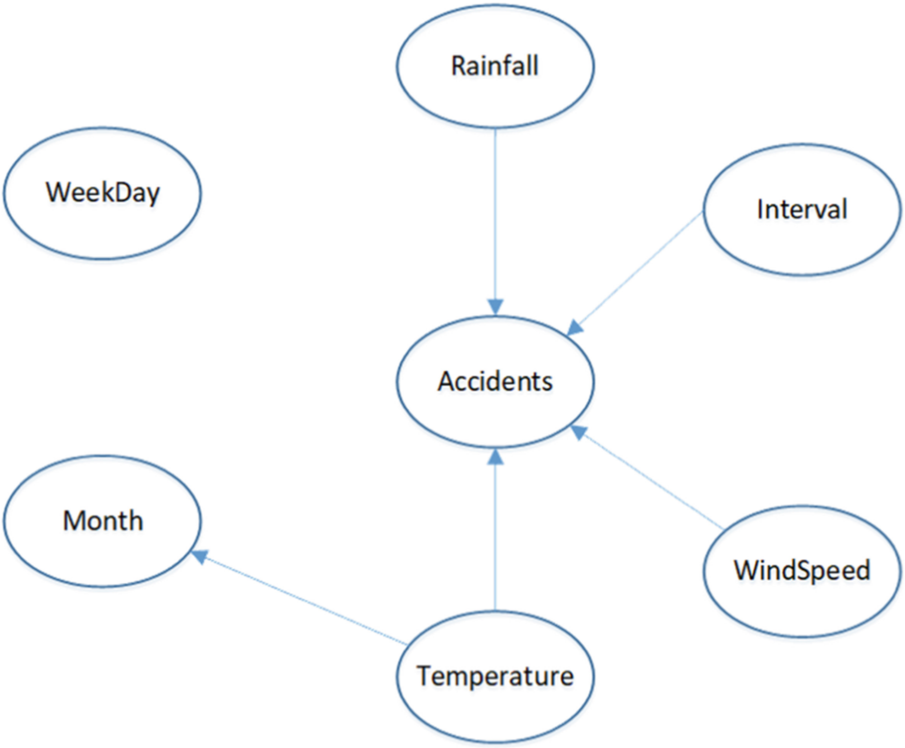
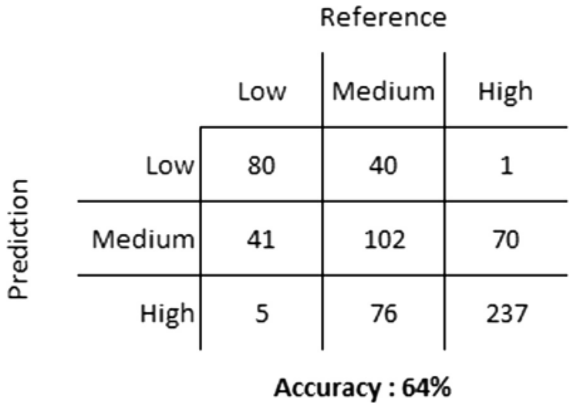


Fig. 4. The Bayesian network obtained by the use of the MuG approach

The results in terms of Confusion Matrix are shown in the following Tables 2, 3 and 4.

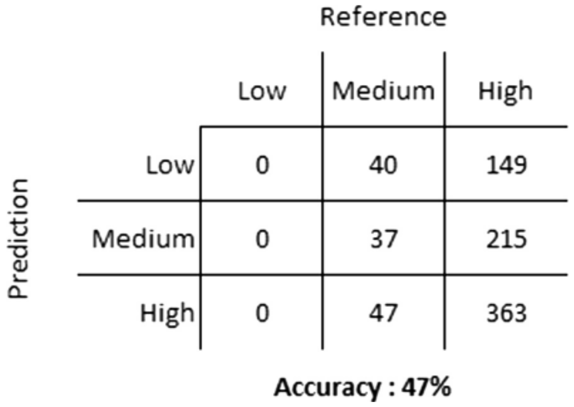
Table 2. Confusion matrix defined Bayesian network



		Reference		
		Low	Medium	High
Prediction	Low	80	40	1
	Medium	41	102	70
	High	5	76	237

Accuracy : 64%

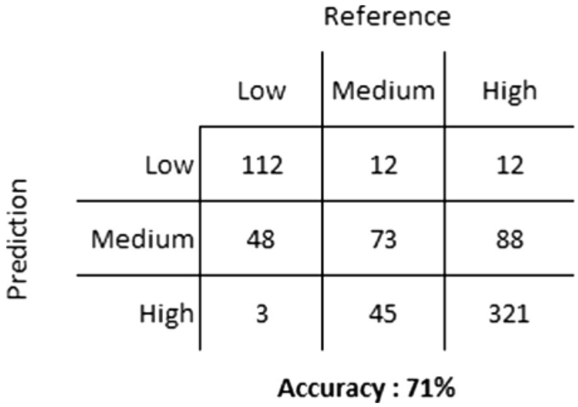
Table 3. Confusion matrix learned Bayesian network



		Reference		
		Low	Medium	High
Prediction	Low	0	40	149
	Medium	0	37	215
	High	0	47	363

Accuracy : 47%

Table 4. Confusion matrix using the MuG approach



		Reference		
		Low	Medium	High
Prediction	Low	112	12	12
	Medium	48	73	88
	High	3	45	321

Accuracy : 71%

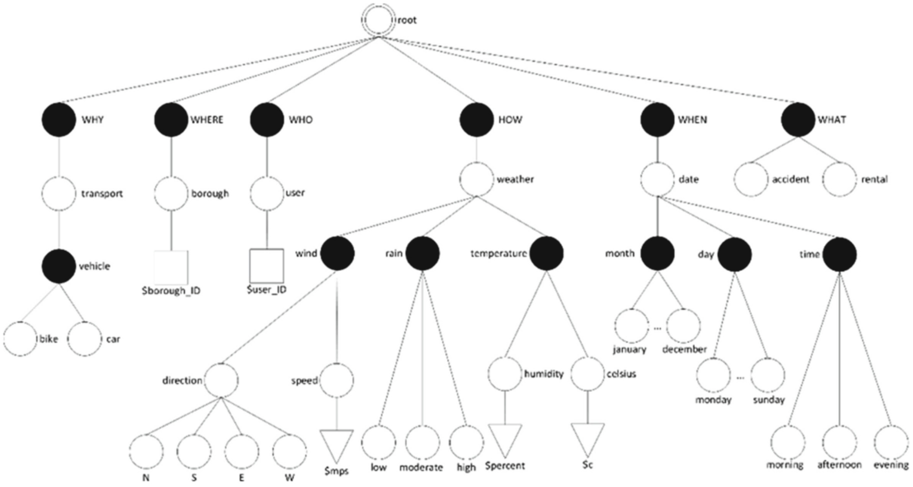


Fig. 5. CDT view

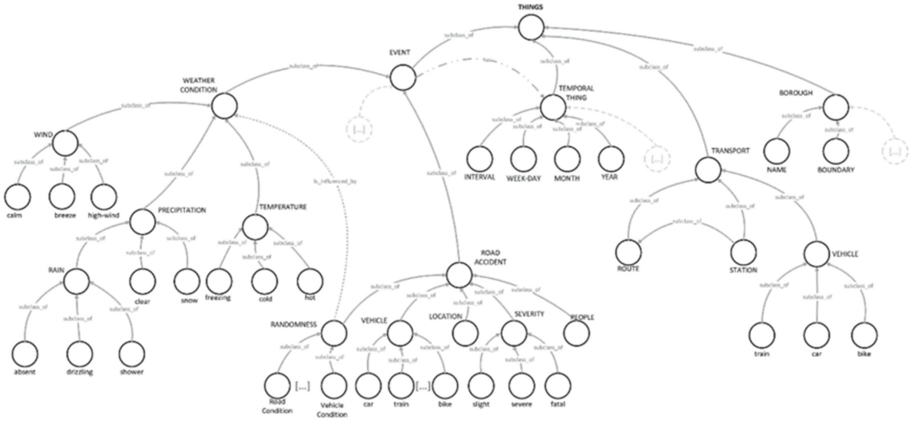


Fig. 6. Ontology view

5 Conclusion

Although the learned Bayesian Network structure by our system is not complete, its reliability in terms of Accuracy increase between Case #1 and Case #3, as shown in the tables. Therefore, we could argue that our system, which is ready to provide reliable answers, can raise its performance over time with increasing volumes of data. Moreover, the potential of such built systems lies in the fact that it is able to automatically update and adapt itself. In addition, it is capable, through Ontologies and CDT, to interface with other similar systems based on different contexts sharing and exchanging knowledge in order to improve its performance more and more.

References

1. deRoos, D., Eaton, C., Lapis, G., Zikopoulos, P., Deutsch, T.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill, New York City (2011)
2. Boyd, D., Crawford, K.: Critical questions for BIG DATA. Inf. Commun. Soc. **15**(5), 1–18 (2012)
3. Laney, D.: 3D Data management: controlling data volume, velocity, and variety. META Group (2001)
4. Tole, A.A.: Big data challenges. Database Syst. J. **4**(3), 31–40 (2013)
5. Colace, F., De Santo, M., Moscato, V., Picariello, A., Schreiber, F.A., Tanca, L.: Data Management in Pervasive Systems. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20062-0
6. Ashton, K.: That "Internet of Things" thing. RFiD J. **22**, 97–114 (2009)
7. Atzori, L., Iera, A., Morabito, G.: The Internet of Things: a survey. Comput. Netw. **54**(15), 2787–2805 (2010)
8. Demchenko, Y., Grosso, P., De Laat, C., Membrey, P.: Addressing big data issues in scientific data infrastructure. In: 2013 International Conference on Collaboration Technologies and Systems (CTS), pp. 48–55 (2013)
9. Harrison, C., Eckman, B., Hamilton, R., Hartswick, P.: Foundations for smarter cities. IBM J. Res. Dev. **54**(4), 350–365 (2010)
10. Zanella, A., Bui, N., Castellani, A.P., Vangelista, L., Zorz, M.: Internet of Things for smart cities. IEEE IoT J. **1**(1), 22–32 (2014)
11. Weber, P., Medina-Oliva, G., Simon, C., Iung, B.: Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. Eng. Appl. Artif. Intell. **25**(4), 671–682 (2012)
12. Colace, F., Lemma, S., Lombardi, M., Pascale, F.: A context aware approach for promoting tourism events: the case of artist's lights in Salerno. Paper presented at the ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems, vol. 2, pp. 752–759 (2017)
13. Panigati, E., Rauseo, A., Schreiber, F.A., Tanca, L.: Aspects of pervasive information management: an account of the green move system In: 2012 IEEE 15th International Conference on Computational Science and Engineering (CSE), pp. 648–655 (2012)
14. Chen, H., Finin, T., Joshi, A.: An ontology for context-aware pervasive computing environments. Knowl. Eng. Rev. **18**(3), 197–207 (2003)
15. Helsper, E.M., Van Der Gaag, L.C.: Building Bayesian networks through ontologies. In: Proceedings of the 15th Eureopean Conference on Artificial Intelligence, ECAI 2002, Lyon, France, July 2002
16. Colace, F., De Santo, M.: Ontology for E-learning: a Bayesian approach. IEEE Trans. Educ. **53**(2), 223–233 (2010)
17. Clarizia, F., Lemma, S., Lombardi, M., Pascale, F.: An ontological digital storytelling to enrich tourist destinations and attractions with a mobile tailored story. In: Au, M.H.A., Castiglione, A., Choo, K.-K.R., Palmieri, F., Li, K.-C. (eds.) GPC 2017. LNCS, vol. 10232, pp. 567–581. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57186-7_41
18. Clarizia, F., Lemma, S., Lombardi, M., Pascale, F.: A mobile context-aware information system to support tourism events. In: Au, M.H.A., Castiglione, A., Choo, K.-K.R., Palmieri, F., Li, K.-C. (eds.) GPC 2017. LNCS, vol. 10232, pp. 553–566. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57186-7_40

19. R.C. Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013 (2014). ISBN 3-900051-07-0
20. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**(4), 309–347 (1992)
21. Scutari, M.: Learning Bayesian networks with the bnlearn R package. J. Stat. Softw. **35**(3), 1–22 (2010)
22. Tucker, A., Trifonova, N., Maxwell, D., Pinnegar, J., Kenny, A.: Predicting ecosystem responses to changes in fisheries catch, temperature, and primary productivity with a dynamic Bayesian network model. ICES J. Mar. Sci. **73**(10), 1334–1343 (2017)