doi: 10.2478/v10313-012-0024-5
2012 /15

Ontology Building Using Data Mining Techniques

Henrihs Gorskis, Yuri Chizhov, *Riga Technical University*

Abstract **– This paper deals with certain data mining techniques in order to discover their potential for use in automated ontology building. The end goal is the reduction in the time requirement for the construction of any given ontology and necessity for expert consultation. This can be achieved by combining data mining and ontology engineering. The aim of this paper is to take a deeper look at potentially useful data mining techniques for an automated ontology building process, to research related publications in this field and to propose ideas on how to use data mining techniques in ontology building.**

Keywords **– data mining techniques, ontology engineering**

I. Introduction

Ontologies are widely used in the fields of knowledge engineering, artificial intelligence, knowledge management, natural language processing, e-commerce, intelligent information integration, bio-informatics, education, semantic web and others [1, 2]. The term 'ontology' in computer science should not be confused with the philosophical meaning used by Plato and Aristotle. An ontology represents the concepts and the relationship between them for a specialized domain or field of interest. A domain ontology can be viewed as a model of the domain and contains structural and conceptual information about it. Building an ontology is complex work. In order to build an ontology usually a domain expert is required to help declare all domain concepts and the relationship between them. This is done mostly manually aided only by ontology editing software that can help the user only to some degree. The automation of this process is much desired and represents a challenge for the fields of computer and information science. The basis for automated ontology building could be within the techniques of data mining. In the context of an expert system, the sequential steps of the construction and usage can be seen in Fig. 1. The basic steps of ontology building First, raw data needs to be gathered. This data can be either structured or not. The data do inherently describe the given domain; however, they are not structured in a way that gives better understanding of it. Using data mining techniques, the concepts, concept-connections and a general structure of the domain model could be found. The result of the applied data mining techniques can then be stored as an ontology.

II. Basic Concepts of Data Mining and Ontology

To better understand the connection between data mining and ontology building it is necessary to elaborate the basic concepts of data mining, the different data mining techniques and the ontology itself. In the given descriptions, we also hint at how the techniques can help in ontology building.

A. Data Mining

Data mining, also called knowledge discovery in databases, in computer science, is the process of discovering implicit, previously unknown, and potentially useful patterns and relationships in large and possibly continuous volumes of data [3]. The objective of data mining is to find ways to automatically detect regularities or patterns in databases [4] for further use. Useful patterns, if found, should be generalized to make as accurate as possible predictions on future data. Thus, the final objective of data mining activity is knowledge discovery. The technical basis of data mining is provided by machine learning. It is used to extract information from the raw data. Abstraction is used in order to find patterns. Usually, it is also necessary that the system provides an explicit structural description, so that to provide the observer with an explanation of what has been learned and an explanation of the basis for new predictions. There are several major data mining techniques, including association, classification, clustering and prediction [5].

B. Clustering

The clustering technique is based on similarities in data objects and the placement of the closest ones into a common cluster. The goal of clustering is the unsupervised detection of meaningful classes for a given set of data objects. Clustering is unsupervised in the sense that there are no previously given target classes and therefore no teacher that controls the

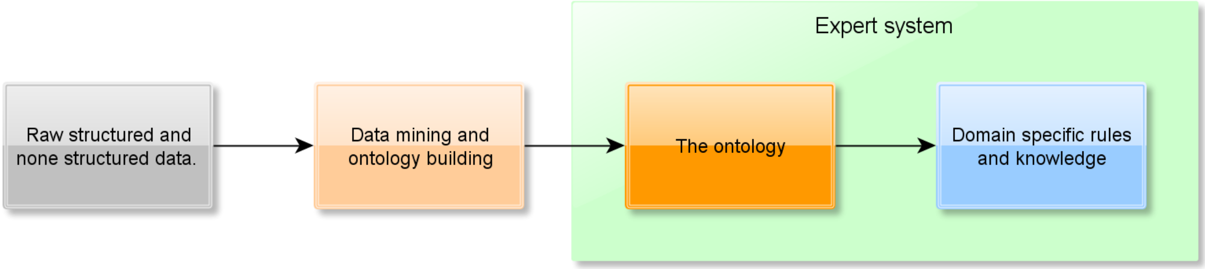
Expert system
Raw structured and none structured data.
Data mining and ontology building
The ontology
Domain specific rules and knowledge

Fig. 1. The basic steps of ontology building

learning process. Clustering can be used for knowledge extraction, data compression and the analysis of the properties of data. Clustering techniques rely on the existence of some suitable similarity metric for objects [4].

When creating an ontology in an unsupervised way (without a domain expert), the technique of clustering can find and provide a base set of classes found within the data. However, some clustering algorithms are distance-based and describe clusters purely by enumerating their members or another arbitrary way. Such algorithms do not provide data useful for naming the detected clusters. Other clustering algorithms represent the clusters by means of a description. This description may take the form of a set of necessary and sufficient conditions for data object to become a member of a given cluster. This set of requirements could provide a basis for naming the cluster, but would require additional methods for doing so, without a guarantee for finding a meaningful name for human interaction. Still, other clustering algorithms use probabilistic descriptions. Probability data can be useful for defining connections between concepts in the ontology. Furthermore, the found set of clusters may be “flat”. It means that no cluster is “contained” in any other cluster. Clusters can also be hierarchical, providing the taxonomy of clusters with definite relationships between them.

C. Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item, in a set of data, into a predefined set of classes or groups, by considering a vector of properties. Different from the clustering technique, classification does not define classes; it only assigns objects to predefined classes. Because of the predefined classes, it is said to be supervised. The classification method uses mathematical techniques such as decision trees, linear programming, neural network and statistics. Usually classification is used in letter recognition, speech recognition, electrocardiogram signal classification and other fields. In the creation of an ontology, classification could be used to teach the ontology a set of rules or parameters by which to assign data to the discovered classes. This could give further information about the properties of the found clusters that serve as class prototypes.

The most common classification algorithm is C4.5. It is used to generate decision trees and is based on the ID3 algorithm.

We suppose that it will also be possible to use this method to create a general ontology structure from a set of domain specific data, but only in the case of the existing predefined classes for the ontology. It is possible to gain an ontology directly from the created decision tree and to use the decision nodes as concept classes and the tree concept structure for the ontology [6].

D. Association Rule Learning

Association rule learning is the discovery of interesting relations in data. Association or associative memory that is addressed by its content can be accessed by specifying the existing content. The content of the memory can be accessed even by partial or distorted input. It means that the content of the associative memory can be accessed even if there is an error in defining the address (input). This technique is one of the best known data mining technique. It discovers patterns based on the relationship of a particular item with other items in the same transaction.

A very popular example of this technique is the market basket analysis, where it identifies what products customers frequently purchase together [7].

This technique has the potential to discover additional and not yet detected connections between concepts for the construction of an ontology. Using statistical data, it is possible to find frequent item sets and to define the probabilistic connections between these item sets to create higher level sets. This forms a structure of classes and super classes. These classes and the structure can be used as the basis for the ontology [8].

E. Prediction

The data mining technique of prediction as it name implied has the task of predicting a certain number of data entries based on the existing data entries. It can discover relationships between independent variables and relationships between dependent and independent variables. It is possible to predict short-term or long-term developments of a given situation. The prediction analysis technique is often used in the sale to predict profit for the future by considering the sale as an independent variable and profit as a dependent variable. Then based on the historical sale and profit data, it is possible to draw a fitted regression curve that is used for profit prediction.

Data mining has two primary goals – prediction and description [6]. The previously mentioned techniques are descriptive and therefore more directly connected to ontology building. Whether the result of prediction can form a direct basis for ontology building is unclear at this point.

Since prediction is used to foresee unknown data, it is feasible that prediction could be used in order to predict data objects that were not given in the existing data. We suppose this could result in better understanding of the model and the relationships between concepts. It also could predict concepts required in the model but could not be found by the given data alone. This could enrich the ontology during its building process.

F. Ontology

An ontology is a specification of a conceptualization, i.e., it is a description of concepts and relationships that exists for an agent or a community of agents [9]. An ontology is a formal explicit description of concepts in a domain of discourse [2]. It consists of classes sometimes called concepts; properties of each concept describing various features and attributes of the concept called slots, roles or properties; and restrictions on slots called facets or role restrictions. The main advantage of ontology-based systems is the ability of communication and sharing knowledge between people and between computer systems [10]. Despite several other applications, an ontology is mainly used to share a common understanding of the structure of information among people or software agents [11].

One could think about it as a kind of dictionary. This gives a common language between the program and the user for the given subject domain. Since it also contains structural and restriction information, it is more than just a dictionary. It enables the reuse of domain knowledge, makes domain assumptions explicit and separates domain knowledge from operational knowledge. It also allows for analyzing domain knowledge [2].

An ontology is basically a model of a domain. The elements of the ontology contain domain specific knowledge and describe the domain as a whole. An ontology can be viewed as a virtual expert of the given domain. Together with the connection to a database that contains facts (facts describe the state of the domain), the ontology that contains information about the domain on a general level and processes that allow obtaining results (interface engine, explanation facilities and a user interface) comprises an expert system. An expert system is a computer system that emulates the decision-making ability of a human expert [12].

It is important to note that we consider the general domain knowledge and rules that are contained within the ontology to be separate from domain state information or facts that can be stored separately in a database. During the ontology building process using data mining techniques, the used data can end up in the ontology as instances of the found classes. We suppose it might be necessary to consider the removal of these instances from the found classes for the final form of the ontology, by keeping only the general class description of the leaf node, if possible, and moving the instance data to the database.

The ontology is the expert knowledge part of an expert system. It must contain all important concepts of the domain, the specific properties of the concepts, the relationships between the concepts (how any given concept affects or is affected by other concepts) and certain additional rules and restrictions of the domain. This information and domain model are general. Within an expert system, it must be possible to retrieve general information about the domain from the ontology on demand. For example, to solve a domain specific planning problem, the expert system will make a call to the ontology to retrieve information on what concepts are connected to the given subject and what rules apply between them. Having retrieved this information, the expert system can connect to the database to obtain current-state information about the obtained concepts and perform calculations or the analysis based on the retrieved rules.

The use, structure and functions of ontologies can be very different; however, the elements contained in ontologies are the same.

The formal definition of any ontology [13] is tuple (1).

$$O = \{C, R, H^C, H^R, I_C, I_R, A\}. \tag{1}$$

The elements of the formal description of the ontology are:

- **C:** represents the set of concepts within the ontology.
- **R:** represents the set of relations between concepts. $R_i \in R$ and $R_i \rightarrow C \times C$.
- **H^C:** represents the hierarchy of concepts in their relations as $H^C \subseteq C \times C$, where $H^C(C_1, C_2)$ means that C_1 is a sub-concept of C_2.
- **H^R:** represents the hierarchy of relations as $H^R \subseteq R \times R$, where $H^R(R_1, R_2)$ means that R_1 is a sub-relation of R_2.
- **I_C:** the set of concept instances.
- **I_R:** the set of relation instances.
- **A^O:** the set of axioms (additional rules and restrictions).

This definition contains all the elements of an ontology; however, it does not define the characteristics of the elements. For example, the concepts within an ontology usually contain a set of attributes.

III. Related Scientific Publications

This paper is based on research of other scientific papers and their proposals and ideas on ontology building and structuring. While getting acquainted with the works of other authors about ontologies and analyzing their approach to the subject, the idea of creating a table for clarity and a better overview has occurred. Table I is the finished overview of the publications and their content.

The first column references the examined work. The full name can be found in the reference section. The information of the second column is based on the formal definition of ontology as given in the previous section and shown in tuple (1). It shows what elements of the definition are mainly focused on and discovered by the methods given in the referenced paper. Some methods yield more elements then others, which in turn define a better ontology. The star that is given for the research paper [14] symbolizes that the goal of the paper has been not to create or find any specific elements of the ontology but to improve the existing ones and the process of ontology building in general. The third column describes the problem domain that the paper has tried to solve. In some cases, the paper has concentrated very much on solving a problem within the domain, in other cases the domain has been used only for context to the ontology engineering problem. The next column gives a general description of the used methods and approaches for solving the given problem. It can be the name of the used algorithm, software or general approach. The fifth column represents the underlying technique used in the method. It tries to give additional information about the method. The sixth column gives an evaluation on the structure and quality of the ontology, since the final structures of the constructed ontologies differ. The last column shows the level of automation possible by the given approach. Semi-automated means that the help of a domain expert or decision-making ontology engineer is still needed. The first three papers listed in the table are the most insightful and give the most information about the automated construction of an ontology.

TABLE I
OVERVIEW OF THE RELATED PUBLICATIONS

Publication	Elements of ontology	Problem domain	Method of solving	Basis of the method	Base structure of the created ontology	Level of automation
[6]	Hc	Building ontology for soybean and animal diseases and their symptoms	C4.5 decision tree algorithm	Classification	Class hierarchy based on decision tree	Fully automated ontology building
[4]	C, Hc	The semantic web and user behaviour ontology building for music recommendation	COBWEB algorithm	Incremental conceptual clustering	Dynamic class hierarchy	Fully automated ontology building
[15]	C, Hc	Ontology building from a text using a multi-agent system	Distributed artificial intelligence, multi-agent system	Distributed hierarchical clustering	Dynamic class hierarchy	Semi-automated
[16]	C, Hc, R, Hr, I	Pharmaceutical tablet production ontology	Manual definition of concept classes and relations	Expert's knowledge	Full ontology	Manual creation by an expert
[17]	C, Hc, R, Hr, I	Combined heart disease, symptom and treatment ontology	Manual elaboration of an existing ontology	Expert's knowledge	Full ontology (the improved existing ontology)	Manual creation by an expert
[12]	*	Improvement of the existing ontology	Knowledge discovery by examining large amounts of uniform data and the usage of the already existing ontology	Combining ontologies and introducing a feedback loop	Full ontology (the improved existing ontology)	Semi-automated
[8]	R, Hc	Neural network for ontology building, example shown on furniture store shopping transactions	Multi-layer feed forward mining	Feed forward neural network	Class hierarchy	Fully automated ontology building
[11]	C, Hc, R, Hr, I	Ontology building for wine classification	Manual definition of concept classes and relations based on but not extending the existing ontology	Expert's knowledge	Full ontology	Manual creation by an expert

It has to be mentioned that the papers that describe a manual approach to ontology building have the most complete and best defined ontology with complex and diverse concept relations and constraints. The same papers also describe the complexity, labour intensity and time consumption that are involved in the manual construction of ontologies, which are far greater than in the automated approaches.

Even though the methods for automated ontology creation construct a more simple structure with more basic relations, the ability of automated ontology creation in unknown or complex problem domains can be very useful. The second paper in Table 1 constructed a complex user behaviour ontology for website users; a domain where experts do not exist.

An example of difference in complexity can be seen in Fig. 2. The manually constructed ontology on the top has usually more relations and of different kinds whereas an automatically created ontology tends to be a tree-like structure (often a binary tree) as seen on the bottom of the image.

IV. APPLYING DATA MINING TECHNIQUES TO ONTOLOGY BUILDING

Some of the mentioned techniques allow using the result of the data mining process directly as the basis for further ontology building or to use it as the ontology itself. However, it is important to note that there seems to be a number of discrepancies between ontology engineering and data mining. Most data mining algorithms that could be used to generate an ontology by creating concepts and/or concept relationships do create a tree graph structure. It is not necessarily a bad thing; however, the definition of ontology does not require a tree graph. An ontology is described as a free form graph of any structure where any concepts can be connected. It would seem that a larger number of relationships would make a better and more descriptive domain model.

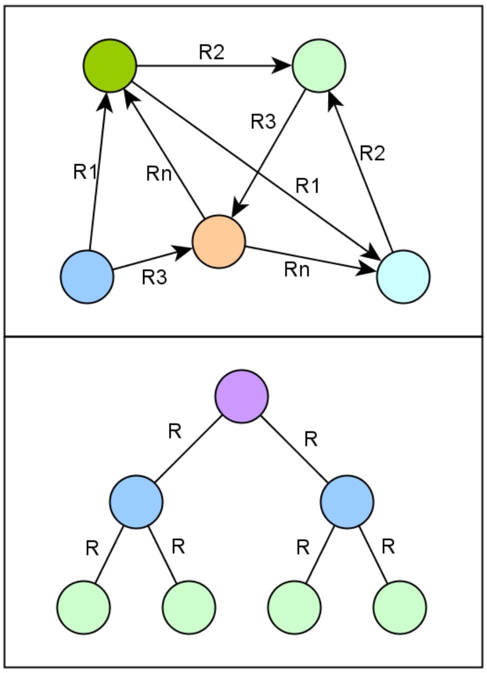

Fig. 2. Ontology structural complexity

This brings us to the second shortcoming of data mining. A generated decision tree or any graph structure obtained from data mining will most likely have only one type of relationship between all connected classes. It is very important for an ontology to have different defined kinds of concept relationships as those contain the domain specific rules. Any complex domain will require diversity in relationships between concepts.

As stated in the description of the classification technique, the C4.5 algorithm creates a decision three. Such a decision tree can be used as the basic structure of the ontology. However, it is not the direct goal of such a tree to find super classes of the given examples, but to construct a structure that will allow for finding the quickest way to choose a leaf node. Such a structure is still usable for ontology building, but this shortcoming needs to be taken into account. Also, since it is based on the supervised classification technique, the examples used for construction need to have a predefined class attribute.

The best algorithm for automated ontology building from a set of data in an unknown domain is the COBWEB algorithm [D]. COBWEB is an incremental conceptual clustering algorithm [D]. It takes the current state of its search, evaluates all the possible steps it can take and chooses the best step for further expansion. As it performs its search, it creates concepts and sub-concepts describing them probabilistically. Every concept found by the algorithm has in its attributes the probabilities for every given example in the dataset. The COBWEB algorithm design was inspired by the conceptual assumption process in human beings.

It is important to define whether or not the finished ontology will also serve as a database for current domain data. When data mining techniques are used for ontology building, the raw data used for learning can end up in the data mining result. If it is desired that the ontology only contains general domain knowledge, these additional data must be removed. If left in a combined state and without additional ontology engineering; the main application of such an ontology could be a database with faster searching capabilities.

V. Conclusions

In this paper, we have reviewed data mining techniques and their potential use in ontology building. All of the techniques can be used for ontology building; classification and clustering are the most useful. The C4.5 and COBWEB algorithms that are based on these techniques are the basis for automated ontology building.

In most cases, the result of the data mining process can be directly stored as an ontology; however, without additional work and/or revision the resulting ontology can in some cases be inferior to manually build ones.

The explored publications shown in Table 1 offer a basis for understanding the methods and techniques used in automated and general ontology building. The papers that describe manual ontology building give a deeper understanding on how an expertly made ontology looks like. The complex relations and concept structures give understanding on what information can be contained within an ontology for use in expert systems. Papers on automated ontology building describe how exactly data mining techniques are used for this task. Even though the final ontology gained from data mining tends to be more simplistic than manually created ones, their use in unknown research domains cannot be understated.

It could be possible in the future to build a general expert system or any kind of computer program without knowing anything about a given field the program is intended for and yet it could perform any domain specific task asked from it, simply by connecting to the specific ontology and retrieving any necessary information on execution from this virtual expert and combining this machine-knowledge with available data.

References

[1] Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho, “Ontological Engineering: With Examples from the Areas of Knowledge Management, E-commerce and the Semantic Web”, 2004, Springer

[2] Natalya F. Noy, Deborah L. McGuinness, “Ontology Development 101: A Guide to Creating Your First Ontology”, 2001, Stanford University, Stanford, CA, 94305

[3] Encyclopedia Britannica, www.britannica.com/EBchecked/topic/1056150/data-mining

[4] Patrick Clerkin, Padraig Cunningham, Conor Hayes, “Ontology Discovery for the Semantic Web Using Hierarchical Clustering”

[5] Data mining techniques, www.dataminingtechniques.net/

[6] Abd-Elrahman Elsayed, Samhaa R. El-Beltagy, Mahmoud Rafea, Osman Hegazy, “Applying data mining for ontology building ”

[7] R Agrawal, T. Imieliński, A. Swami, “Mining association rules between sets of items in large databases”, 1993, Proceedings of the 1993 ACM SIGMOD international conference on Management of data

[8] A. Bhagat, S. Sharma, K. R. Pardasani, “Ontological Frequent Patterns Mining by potential use of Neural Network”, 2011, International Journal of Computer Applications

[9] N. Chalortham, P. Leesawat, M. Buranarch, T. Supnithi, “Ontology Development for Pharmaceutical Tablet Production Expert System”, 2008, Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference

[10] Darya Plinere, Arkady Borisov, “SWRL: Rule Acquisition Using Ontology”, 2009, Scientific Journal of Riga Technical University.

[11] J. Graca, M. Mourao, O. Anunciacao, P. Monteiro, H. S. Pinto, V. Loureiro, “Ontology building process: The wine domain”

[12] Jackson, Peter (1998), Introduction To Expert Systems (3 ed.)

[13] Yildiz Burcu, “Ontology Evolution and Versioning: The state of the art”, 2006, Vienna University of Technology

[14] M. d’Aquin, G. Kronberger, M. C. Suárez-Figueroa, „Combining Data Mining and Ontology Engineering to enrich Ontologies and Linked Data“

[15] K. Ottens, M.-P. Gleizes, P. Glize, „A Multi-Agent System for Building Dynamic Ontologies“, 2007, IFAAMAS

[16] N. Chalortham, P. Leesawat, M. Buranarach, T. Supnithi, “Ontology Development for Pharmaceutical Tablet Production Expert System”, 2008, Proceedings of ECTI-CON

[17] A. Jovic, D. Gamberger, G. Krstacic, “Heart failure ontology”

Henrihs Gorskis is a second year graduate student of Riga Technical University (RTU). He received his BSc.ing. in 2011. His research interests include data mining, ontology engineering, evolutionary computing and programming. E-mail: henrihs.g@gmail.com

Yuri Chizhov is currently a Lecturer and Senior Laboratory Assistant at the Department of Modelling and Simulation, Riga Technical University (RTU). He received his Dr.sc.ing. from Riga Technical University in 2012. His major research interests include a number of techniques related to computational intelligence, in particular cluster analysis, ontology building. E-mail: jurij.ch@gmail.com

Henrihs Gorskis, Jurijs Čižovs. Ontoloģijas būvēšana, izmantojot datu ieguves metodes
Ontoloģija ir modelis, kas izmanto konceptus un saites starp tiem, lai aprakstītu noteiktu nozari. Tas tiek izmantota, lai definētu kopīgu terminoloģiju un nozares zināšanas, kas ir kopīgi pieejamas lietotājam un datorprogrammām. Ontoloģijas uzbūve ir laikietilpīgs un sarežģīts process, kurā parasti piedalās nozares eksperts. Lai atvieglotu un paātrinātu šo procesu, kā arī lai atbrīvotos no nepieciešamības iesaistīt ekspertu, tiek piedāvāts izmantot datu ieguves tehnikas ontoloģiju būvēšanas automatizācijai. Tika apskatīti klasterizācijas, klasifikācijas, asociācijas un prognozēšanas paņēmieni un izvirzīti priekšlikumi to lietošanai automatizētā ontoloģijas būvēšanā. Tika dota ontoloģijas formālā definīcija, kurā tiek nosaukti visi tās elementi. Tika veikts saistīto zinātnisko publikāciju apskats tabulas veidā, kurā ir izklāstīts, kādi paņēmieni tika izmantoti un kādi ir iegūtie rezultāti. Ne visi apskatītie darbi piedāvāja ontoloģijas uzbūves automatizācijas iespējas, daži aprakstīja manuālo pieeju. Veiktajā pārskatā atklājās, ka datu ieguves tehnikas tiešām ir lietojamas ontoloģijas uzbūves automatizācijai un var sniegt iespaidīgus rezultātus. Bieži datu iegūšanas algoritmu rezultāts tiešā veidā tiek pārņemts ontoloģijā. No citu darbiem tika iegūts, ka manuāli būvētas ontoloģijas tomēr mēdz būt dziļāk aprakstītas un satur vairāk informācijas. Automātiski būvētas ontoloģijas dotajā brīdī ir vienkāršākas, tomēr tiek veidotas daudz ātrāk un bez nozares eksperta. Tas ļauj izmantot automātisku ontoloģijas uzbūvi tādās nozarēs, kurās nav ekspertu, kā arī atklāt un iegūt līdz šim nezināmo informāciju. Vispiemērotākais algoritms dotajam uzdevumam ir ar nosaukumu COBWEB, kas izmanto neuzraudzīto klasterizāciju un varbūtības mērus konceptu atklāšanai.

Генрих Горский, Юрий Чижов. Построение онтологий с использованием методов интеллектуального анализа данных
Онтология представляет собой модель, которая использует понятия и связи между ними, чтобы описать определённую отрасль. Она используется, чтобы определить общую терминологию и знания об отрасли, которые общедоступны для пользователя и программного обеспечения. Построение онтологий является длительным и сложным процессом, в котором, как правило, участвует отраслевой эксперт. Для облегчения и ускорения этого процесса, а также чтобы избавиться от необходимости привлечения экспертов, предлагается использовать методы добычи данных для автоматизации построения онтологий. Были рассмотрены методы кластеризации, классификации, ассоциации и прогноза, а также выдвинуты предложения по их использованию в автоматизированном построении онтологии. Дано формальное определение онтологии, которое обозначает все ее элементы. Был сделан обзор научных публикаций в виде таблицы, в которой излагалось, какие методы использовались и какие были получены результаты. Не все рассмотренные статьи давали возможность автоматизированного проектирования онтологии, некоторые предполагали ручной подход. Проведённый обзор показал, что методы интеллектуального анализа данных используются для автоматизации построения онтологии и могут обеспечить впечатляющие результаты. Часто результат алгоритмов добычи данных непосредственно передаётся в онтологию. Из других статей стало известно, что вручную построенные онтологии, как правило, более глубоко описаны и содержат больше информации. Автоматически построенные онтологии в данный момент являются более простыми, однако они создаются намного быстрее и без отраслевого эксперта. Это позволяет автоматическое построение онтологии в отраслях, где нет экспертов, а также выявление и получение ранее неизвестных сведений. Наиболее подходящим алгоритмом для данной задачи является COBWEB, который использует самообучающуюся кластеризацию и вероятностные меры для обнаружения концептов.