Intelligent Information Management, **2013, 5, 10-17**
http://dx.doi.org/10.4236/iim.2013.51002 Published Online January 2013 (http://www.scirp.org/journal/iim)

A State-of-the-Art Survey on Semantic Web Mining

Qudamah K. Quboa, Mohamad Saraee
School of Computing, Science, and Engineering, University of Salford, Salford, UK
Email: q.k.m.quboa1@edu.salford.ac.uk, m.saraee@salford.ac.uk

Received November 9, 2012; revised December 9, 2012; accepted December 16, 2012

ABSTRACT

The integration of the two fast-developing scientific research areas Semantic Web and Web Mining is known as Semantic Web Mining. The huge increase in the amount of Semantic Web data became a perfect target for many researchers to apply Data Mining techniques on it. This paper gives a detailed state-of-the-art survey of on-going research in this new area. It shows the positive effects of Semantic Web Mining, the obstacles faced by researchers and propose number of approaches to deal with the very complex and heterogeneous information and knowledge which are produced by the technologies of Semantic Web.

Keywords: Web Mining; Semantic Web; Data Mining; Semantic Web Mining

1. Introduction

Semantic Web Mining is an integration of two important scientific areas: Semantic Web and Data Mining [1]. Semantic Web is used to give a meaning to data, creating complex and heterogeneous data structure, while Data Mining are used to extract interesting patterns from, homogenous and less complex, data. Because of the rapid increasing in the amount of stored semantic data and knowledge in various areas, as the case in biomedical and clinical scenarios, this could be transformed to a perfect target to be mined [2,3] leading to the introduction of the term "Semantic Web Mining". This paper gives a general overview of the Semantic Web, and Data Mining followed by an introduction and a comprehensive survey in the area of Semantic Web Mining.

2. Semantic Web

The Semantic Web is changing the way how scientific data are collected, deposited, and analyzed [4]. In this section, a short description defining the Semantic Web is presented followed by the reasons behind the developing of Semantic Web. Next a few selective representation techniques recommended by W3C are presented and a number of successful examples from the commercial domain that support and use the semantic data are given as well.

2.1. Semantic Web: Definition

Semantic Web is about providing meaning to the data from different kinds of web resources to allow the machine to interpret and understand these enriched data to precisely answer and satisfy the web users' requests [1,5, 6]. Semantic Web is a part of the second generation web (Web2.0) and its original idea derived from the vision W3C's director and the WWW founder, Sir Tim Berners-Lee. According to [5] Semantic Web represents the extension of the World Wide Web that gives users of Web the ability to share their data beyond all the hidden barriers and the limitation of programs and websites using the meaning of the web.

2.2. Reasons behind Developing Semantic Web

As noted by [7], the Semantic Web is introduced to crack two specific problems: the limitations of data access in the web (for example retrieving documents according to a given request using ambiguous terms, and the current retrieving systems problem of acquiring only a single "best fit" documents for a query), and the delegation tasks' problems (such as integrating information) by supporting access to data at web-scale and enabling the delegation of certain classes of tasks.

2.3. Semantic Web Representation Techniques

Many available techniques and models are used to represent and express the semantic of data such as the standard techniques recommended by W3C named Extensible Markup Language, Resource Description Framework, and Web Ontology Language [5] which are briefly explained below.

2.3.1. Extensible Markup Language

The Extensible Markup Language (XML) technique has

been established as a generic technique to store, organize, and retrieve data on/from the web. By enabling users to create their own tags, it allows them to define their content easily. Therefore, the data and its semantic relationships can be represented [7,8].

2.3.2. Resource Description Framework

The Resource Description Framework (RDF) is a common language that enables the facility to store resources' information that are available in the World Wide Web using their own domain vocabularies [5,6]. Three types of elements contented in the RDF: resources (entities identified by Uniform Resource Identifiers URIs), literals (atomics values like strings and numbers), and properties (binary relationships identified by URIs) [3]. This is a very effective way to represent any kind of data that could be defined on the web [5].

2.3.3. Web Ontology Language

The Web Ontology Language (OWL) is considered a more complex language with better machine-interpretability than RDF. It precisely identifies the resources' nature and their relationships [8]. To represent the Semantic Web information, this language uses ontology, a shared machine-readable representation of formal explicit description of common conceptualization and the fundamental key of Semantic Web Mining [6,8]. Ontology creators are expressing the interest domain which is based on classes, and properties (represent atomic distinct concepts and rules in other semantic languages respectively) [9].

As examined by [10], the architecture of Semantic Web that is based on the vision of Sir Berners-Lee, is divided into seven layers: 1) URI; 2) XML, NS, & XML schema; 3) RDF & RDF schema; 4) the ontology vocabulary; 5) Logic; 6) Proof; and 7) Trust.

First of all, URI which is in charge of resource encoding process and its identification. Next, XML, NS, and XML schema layer which is in charge of 1) the separation of data content, data structure, and the performance format based on linguistic; and 2) representing them using a standard format language. Furthermore, the layer of RDF and RDF schema define the information on World Wide Web and its type using a semantic model. Moreover, the ontology vocabulary layer is concentrated on revealing semantics among information by defining the knowledge shared and the semantic relations within different sorts of information. Logic is the next layer which takes the responsibility of providing the foundation of intelligent services such as logical reasoning by supplying axioms and inference principles. The last two layers are "proof" and "trust" which deal with enhancing the security of web by using encryption and digital signature mechanisms to identify changes in documents. This architecture of Semantic Web is shown in **Figure 1**.

2.4. Semantic Web in the Commercial Domain

It has been noticed by [7] that there has been an argument about the limitation of applying Semantic Web in the commercial domain and it is restricted to the educational domain only. This claim was overruled by the announcements of trademark companies and the production of commercial products and applications such as Oracle 11 g system which support a large number of core technologies including RDF and OWL. The new apple application is named SIRI which is a virtual personal assistant for a number of general tasks using the Semantic Web services as shown in **Figure 2**.

3. Data Mining

This part is introducing the concept of data mining followed by description of few popular data mining technique and a brief overview of Web Mining and its types.

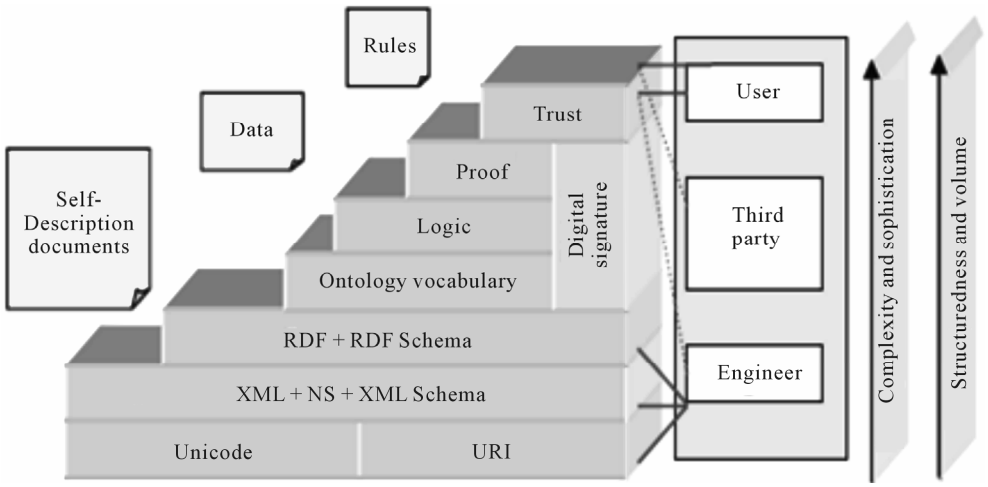

Figure 1. The semantic web layers architecture [6].

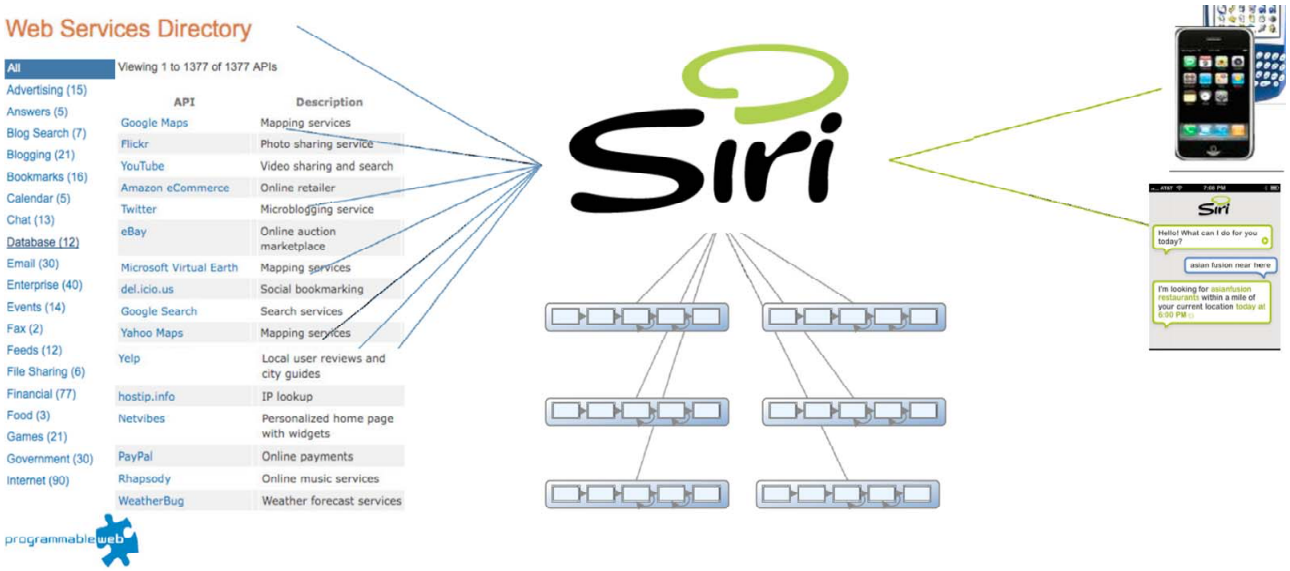

Figure 2. The main architecture of SIRI [7].

Data mining has a great role in the development of many domains especially the educational one. It can be described as a way to find hidden, novel, formerly undiscovered, interesting patterns and rules in huge data sets by using complex tools such as statistical methods, mathematical models, and machine learning algorithms [8].

One of the widely used techniques of data mining is Association Rules Mining which could be defined as algorithm of generating patterns in the form of X => Y, each antecedent (X) and consequent (Y) are non-empty subset of items, and $X \bigcap Y = \varnothing$ and the rule has a support and confidence values higher than the users' thresholds [3,9]. This could be done through two steps: find out frequent subsets of items that satisfy user's support thresholds, then produce patterns with regard to the user's confidence level and based on the frequent subsets. Association rules mining could be categorized into several types for example: either in accordance with the cardinality between X and Y, such as horn-like rules, etc., or in accordance with what is based on the selection criteria of the item sets used in the process of generating rules, like generalized rules, quantitative association rules [3].

It has been reported by [8] that Web Mining is a combination of both text and data mining to mine the biggest information resource (Web). Using data mining techniques in the web is a useful method to extract potentially valuable information from documents in the web. Reference [1] observes that Web Mining could be divided into: Web Usage Mining (WUM), Web Content Mining (WCM), and Web Structured Mining (WSM).

First of all, WUM is focused on discovering what users of web are searching for [1]. This is done by mining web user's log records to possibly gain interesting information about the use of browser and page links that could be helpful in understanding users' behaviour, leading to personalize user's service [10]. Next, WSM, which is a technique, tries to recognize the topology information of network, mining connections between web information pages [10]. Finally, WCM is concentrating on mining information and knowledge from different contents such as text and image [10] attempting to recognize the patterns of specific web users based on their interest or according to their specific regions [1] to help and improve results of many areas like search engines to deliver more precise and beneficial information to web users [10].

4. Semantic Web Mining

This section provides a more explained introduction to the Semantic Web Mining followed by few examined problems facing mining the semantic data with their possible solutions (proposed by researchers) and then selective cases examined where obstacles faced traditional, data mining, and Semantic Web systems (and applications), where using the Semantic Web Mining could possibly help to tackle them and proving its usefulness in different domains. A summary of the reviewed research papers is provided at the end.

4.1. Semantic Web Mining Definition

The huge growing in the quantity of semantic data and knowledge in different fields, as the circumstance in biomedical and clinical scenarios, could possibly create a perfect and important target in the mining process [2,3].

The Semantic Web Mining came from combining two interesting fields: the Semantic Web and the data mining [1]. A possible architecture of this kind of mining suggested by [3] is described in **Figure 3**.

Mining Semantic Web ontologies provides a great possibility to get better results to its domain [3,11], discovers new and valuable insights data from the semantic annotations [12], solves problems that deals with complex and heterogeneous data [3,9] and improves in easy, and effective ways the results of the web mining [10,13].

There is a need to apply and adapt data mining techniques to extract information and knowledge efficiently and effectively, represented in Semantic Web data, and to enhance the way these data are used. The requirement for a shift mining data to mining of semantic data came from adoption of Semantic Web concepts and representations in many different areas such as communities, blogs, search engines and portal, leading to fast growth in the amount of semantic data as shown in **Table 1** which shows statistical results from Falcons and Swoogle Semantic Web search engines [5].

The Semantic Web portal service provided by Twine is another example which reveals this need. Twine saves users' information and interest using RDF and OWL; Twine has more than three millions semantic tags and millions of relations [5].

4.2. Semantic Web Mining: Challenges

When applying Semantic Web Mining, several possible challenges will appear because of different issues such as the complexity and the nature of the semantic data [3] and few possible problems with their suggested solution(s) are described below.

Reference [3] stated that one of the main obstacles in mining the Semantic Web data is recognizing interesting transactions and items from the semi-structured data and that could be caused by three reasons: firstly, the traditional data mining algorithms are built to mine homogeneous data sets. Secondly, the normal way of representing the semantic data is by triple structure consisting of subject, predicate, and object (SPO) and each triple defines a fact which causes the complexity in the data. Finally, most sublanguages of OWL are provided by description logics, "*knowledge representation formalisms with well-understood formal properties and semantics*" [3], instances from the same OWL class might have multiple structures causing the heterogeneous nature of the data. Different solutions are proposed to overcome these difficulties, for example handling the hidden knowledge in semantic data by applying a kind of semantic reasoner, and a pre-process of the triples is done by calculating the composition values followed by grouping and then constructing transactions under specific considerations according to the user's requirement. The resulting paper is very well organized, has a clear methodology and contains all the required and relevant information, but the

Table 1. The amount of semantic web data adopted from [5].

Selected dates	Falcon search engine statistics	
	RDF/XML	Quadruple
09/09/02	21,639,337	2,936,868,638
09/05/29	19,919,364	2,177,084,709
Selected dates	Swoogle search engine statistics	
	Semantic web document	Triple
09/10/17	3,109,616	1,065,799,526

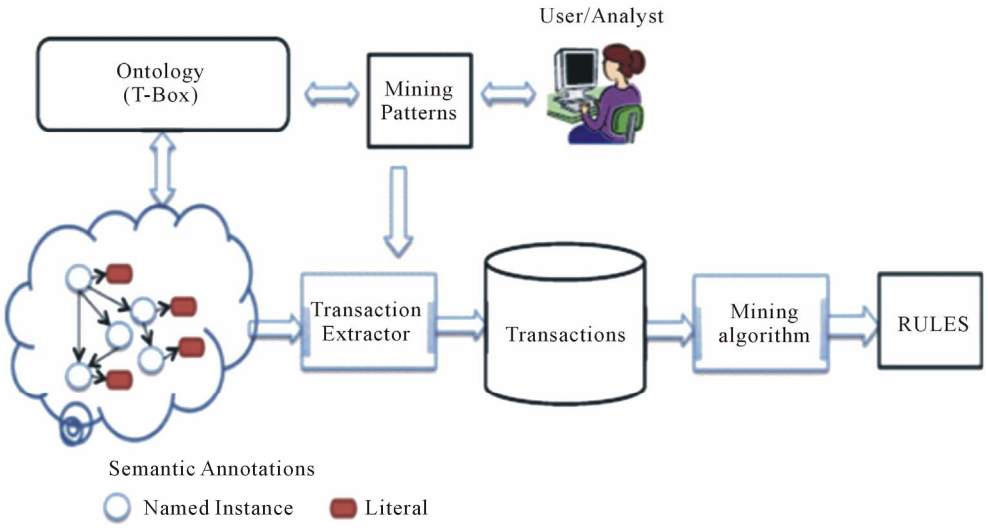

Figure 3. Semantic web mining architecture example [3].

results show that the generated rules have low level of support. More work on increasing the support values and the acceptability of generated results is therefore required. The used dataset, from a biomedical domain, is very reliable and its explanation is very clear and its total number of semantic annotations could be considered as appropriate sample size.

Reference [5] argued that a problem of Semantic Web ontology structure appears when a traditional decision tree algorithm is trying to make practical use of extra information from ontology, and when this mining algorithm is trying to select variables correctly and that because of the network composition of the ontologies in the semantic data leading to the possibility of unlimited number of properties (no restriction) and each property is allowed to content multiple values. Therefore, a number of modifications are proposed to overcome these limits such as including information about relations between concepts and roles (named properties in OWL) of objects based on ontology in the mining process, using description logic based constructor to increase the power of condition's expression and providing a method for choosing variables automatically using statistical basis and ontology relations' information.

In mining semantic data, the problem of non-frequent items could happen leading to reduce the support thresholds of the mining process which could produce a considerable number of irrelevant patterns. A possible answer to deal with this obstacle is generalizing the values of items (objects or subjects) based on the kinds of the concepts and the concepts themselves [12]. In their work, it appears that the use of Semantic Web data is limited by using only two of the triple data structure in the mining. However, there is an absence of relevant information about the dataset sample being used, such as the sample size, and the sample annotations. There is a lack of evaluation and validation of the proposed techniques, and the conclusion is very general and does not describe the presented work, with weak justifications.

4.3. Semantic Web Mining: Cases to Support Its Usefulness

On the users' intention search, [9] explains that a possible problem could happen when a user is searching about specific ontology concepts (set of features) using a learning algorithm. The selected features may negatively affect the items' selections which form a transaction related to the appearing of the conceptual entities under different context in an ontology causing an ambiguity problem. This challenge of correctly interpreting the users' intentions in the system automatically and providing the correct query, a possible solution suggested by the researchers is to provide the user with the ability to select the required context with a keyword which is attached to the suited concept and developing the system to handle the searching for appropriate contexts. Using this kind of restrictions could decrease the searching space of features, and reduce the number of uninterested transactions which means producing less complexity and overload. This work is implemented in a proper and organized way. The dataset is very dependable and clarified very well. The sample size is very suitable for mining purposes. However, the presented approach possibly needs more effort to gain better results (increasing the low level of the generated patterns' support is required).

According to [2], the incorporating with knowledge domain is commonly mentioned as one of the very serious obstacles facing the data mining technologies. One of the reasons behind this problem could be that the knowledge representations are used to be coded under specific applications' formats (scope and granularity) and that causes many difficulties in mining the knowledge. The Semantic Web technologies started to be used to model and store domain knowledge leading to create huge amount of semantic data, and shifting the existing data mining community's paradigm to the new mining technology (Semantic Web Mining) to incorporate with knowledge domain. Their work shows shortcomings in the testing, evaluation and validation of the proposed algorithm. There is a total absence in the dataset's information, and used sampling methodology, which has led to the (probably) unreliable results. The conclusion requires more support and validation.

The work presented by [4] is supported the requirement of the paradigm shifting from traditional mining of abundant empirical data reinforced by knowledge to mining the abundant knowledge data created in the domain of ontologies, supported by rules and models built from heuristics, and statistical computation from data collection. On one hand, Semantic Web Mining could give more support to the users by delivering semantically meaningful results. On the other hand, this possibly allows for developing a great utility that provides improvements in the mining results especially in domains such as biomedical, finance, sociology, and biology where data are mixed with semantic descriptors (represented by ontologies). The study is well explained and organized. The datasets used are very good samples and supported with good explanations. Although it demonstrates the effectiveness of the suggested prototype, the implementation of the proposed algorithms requires more clarification especially SEGS.

A case study presented by [13], mentions a problem that faces users of the web while they are navigating and retrieving information from websites. Most of the time the web users have obstacles to get their requirements causing dissatisfaction and turn away from this website.

The vendor, consequently lose valuable customers and because of the later intention to satisfy and grasp as much as possible of customers, they require a high level of accuracy and correctness of the provided web content. Most of the existing applications suffer from the limitation of retrieved information from content not in a text format such as flash animations, videos, and images causing the analysis process to be done on a small proportion of the whole data. The proposed solution is to use Semantic Web Mining to discover a novel relation from the components' structures of a WebObjects, "*a structured group of words or a multimedia files present within a Webpage that has metadata for describing its content*" [13], based on web user's perspective and possibly leading to enhance the website's satisfaction level by empowering the information provided and considering its preferred appealing format according to the preferences of web user. The presented research shows great intention in the details and is well structured. However, the use of a very small size of participants, only 10 users, to test, evaluate, and validate the proposed algorithm could lead to incorrect and unreliable results. There is information missing about the methodology used to select the representative sample, and the experiment's details such as where and how the experiment was conducted.

The work presented by [2], examines the need for more powerful automatic suggestions systems especially after the vast increase in the use of Semantic Web ontologies on the web. Most of existing Semantic Web search systems are causing number of hidden problems for users of web in selecting appropriate Semantic Web features and terms since this task required to be acquainted with the defined semantic ontologies which could be solved by a learning-based semantic search using Semantic Web Mining technique to combine different measurement techniques such as conceptual comparisons and structural similarities to decide the match degree of a document compared to the user's searched terms. The proposed system recommends proper terms and features (ontologies) for annotation by providing related information, related keywords, and domain information semantically. As mentioned previously, this study requires more work on the evaluation and validation of the proposed algorithm, it has poor explanation for the used dataset, and the conclusion is not well supported.

In regard to the accuracy of retrieving information from the web, [8] examined the effect of applying Semantic Web Mining technique to semantic data in the educational domain, in particular in the distance learning system (e-learning). Their test shows positive effects in the use of relationships and logics among ontologies in expanding system query's search and increasing system's capability to convey web users' willingness and possibly refining and improving the accuracy and recall percentage of retrieved information from web in a semantic search system. This modification in the way systems' search could move these systems to upper level in their functionality precision, the quality services provided, and the interoperability with the educational field standards. The work is supported by a good validation but the dataset created requires more details. Also, the definition provided of Semantic Web Mining does not accurately reflect the correct definition, since it explains the Semantic Web architecture.

Even after using the Semantic Web in the e-learning field, the e-learning is still limited because of the very important and known obstacle of the communication between both the tutors and students, and students and their advisors. This obstacle is happening since all the information and material uploaded and accessed using the web without face to face contact compared to traditional learning system. This limitation is causing problems in tracking students' situations, giving proper instructions to improve their performance, etc. To reduce this gap between the two learning systems, Semantic Web Mining proposed to investigate students' logs data on distance learning portals to provide signs, information about students' conditions and what could motivate and help them, to the administrators and advisors to decide the best way to guide their students to more successful study and by personalizing of e-learning content and services provided according to each student's preferred studying strategies [1]. From their work, it appears that the representation of the semantic data, collected by questionnaire, using a relational database is not the best way, since there is a more suitable format such as XML, RDF, and OWL which shows the real semantic data representation. Since a normal relational database has been used, it seems that this is inappropriate Semantic Web Mining.

A summary of selective works related to Semantic Web Mining is given in **Table 2**.

5. Conclusion

Semantic Web Mining is a new and fast-developing research area combining Web Mining and Semantic Web. In this paper a detailed state-of-the-art survey of on-going research in Semantic Web Mining has been presented. This study analyzes the merging of trends from both areas including a) using semantic structures in the Web to enrich the results of Web Mining and b) to build the Semantic Web by employing the Web Mining techniques. We also have provided justification that the two areas Web Mining and Semantic Web need each other to achieve their goals, but that the full potential of this convergence is not yet realized.

Table 2. Summary of selected existing works in semantic web mining.

Reference number	Existing semantic web mining works' summary			
	Domain	Dataset	Mining technique(s)	Results
[1]	Distance learning	Data from index of learning styles questionnaire	Association rules (Apriori algorithm)	The results are dependable and statistically reliable and significant.
[2]	Knowledge		Weighted feature-based search model	
[3]	Biomedical	Biomedical semantic data and data from health-e-child project	Association rules	Useful approach, promising results, and self-explained rules.
[4]	Biology	Simple handcraft scenario (Bank) and two functional genomics scenarios	Classification (CN2, CN2-SD, SEGS & g-SEGS)	Better than traditional systems based on generalization of rules and automatically pre-process of data.
[5]		Person data, train heading	Semantic decision tree	The presented algorithm has more complex and rich expression, has a number of limitations.
[8]	Education	Data from students, faculties and courses.	Association rules	The presented method improves the precision and recall metrics of web retrieval system.
[9]	Medical	Data from health-e-child project	Association rules (Apriori)	Promising results (low support), useful method.
[10]	Agents		Clustering	Further study is required.
[12]		DBpedia data set	Association rules	Further research is required.
[13]	Web personalisation	Chilean geographical information systems service provider website (named DMapas website)	Clustering (SOFM and K-mean algorithms)	The methodology proposed proves its effectiveness with 80% as a minimum.

6. Acknowledgements

Special thanks to the Higher Committee for Education Development in Iraq (HCED Iraq) to support this research.

REFERENCES

[1] O. Mustapaşa, A. Karahoca, D. Karahoca and H. Uzunboylu, "Hello World, Web Mining for E-Learning," *Procedia Computer Science*, Vol. 3, No. 2, 2011, pp. 1381-1387. doi:10.1016/j.procs.2011.01.019

[2] H. Liu, "Towards Semantic Data Mining," *Proceedings of the 9th International Semantic Web Conference*, Shanghai, 7-11 November 2010, pp. 1-8.

[3] V. Nebot and R. Berlanga, "Finding Association Rules in Semantic Web Data," *Knowledge-Based Systems*, Vol. 25, No. 1, 2012, pp. 51-62. doi:10.1016/j.knosys.2011.05.009

[4] N. Lavrač, A. Vavpetič, L. Soldatova, I. Trajkovski and P. K. Novak, "Using Ontologies in Semantic Data Mining with SEGS and G-SEGS," *Proceedings of the 14th International Conference on Discovery Science*, Espoo, 5-7 October 2011, pp. 165-178.

[5] D. Jeon and W. Kim, "Development of Semantic Decision Tree," *Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications*, Macau, 24-26 October 2011, pp. 28-34.

[6] V. Sugumaran and J. A. Gulla, "Applied Semantic Web Technologies," Taylor & Francis Group, Boca Raton, 2012.

[7] J. Domingue, D. Fensel and J. A. Hendler, "Handbook of Semantic Web Technologies," Springer-Verlag, Heidelberg, 2011.

[8] A. Jain, I. Khan and B. Verma, "Secure and Intelligent Decision Making in Semantic Web Mining," *International Journal of Computer Applications*, Vol. 15, No. 7, 2011, pp. 14-18. doi:10.5120/1962-2625

[9] V. Nebot and R. Berlanga, "Mining Association Rules from Semantic Web Data," *Proceedings of the 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, Córdoba, 1-4 June 2010, pp. 504-513.

[10] W. Yong-Gui and J. Zhen, "Research on Semantic Web Mining," *Proceedings of the International Conference on Computer Design and Applications*, Qinhuangdao, 25-27 June 2010, pp. 67-70. doi:10.1109/ICCDA.2010.5541057U

[11] A. Segura, C. Vidal-Castro, V. Menéndez-Domínguez, P. G. Campos and M. Prieto, "Using Data Mining Techniques for Exploring Learning Object Repositories," *The Electronic Library*, Vol. 29, No. 2, 2011, pp. 162-180. doi:10.1108/02640471111125140

[12] Z. Abedjan and F. Naumann, "Context and Target Configurations for Mining RDF Data," *Proceedings of the 1st International Workshop on Search and Mining Entity-Relationship Data*, Glasgow, 24-28 October 2011, pp. 23-24. doi:10.1145/2064988.2064998

[13] J. D. Velásquez, L. E. Dujovne and G. L'Huillier, "Extracting Significant Website Key Objects: A Semantic Web Mining Approach," *Engineering Applications of Artificial Intelligence*, Vol. 24, No. 8, 2011, pp. 1532-1541. doi:10.1016/j.engappai.2011.02.001

