International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

# Survey on Current Trends and Technologies of Semantic Web Mining

**Khin Mya Nwe**

*Department of Information Technology Support and Maintenance, University of Computer Studies, Yangon, Myanmar

**Abstract:** ***Semantic Web is a product of Web 2.0 that is supported with automated semantic agents for processing user data to help the user on ease of use and personalization of services. Web Mining is an application of data mining which focuses on discovering patterns from Web logs and data. The semantic structure can be built with the pattern or relation results discovered via Web Mining. 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. The objective of this paper is to provide an outline of Semantic Web and Web mining technologies and to give a survey on current researches in the area of Semantic Web Mining.***

**Keywords:** Semantic Web; Web Mining; Semantic Web Mining, Resource Description Framework, Ontology

## 1. Introduction

Nowadays, the Web is rapidly growing and becoming a huge repository of information, with several billion pages and more than 300 million of users globally. The nature of most data on the Web is so unstructured that they can only be understood by humans, but the amount of data is so huge that they can only be processed efficiently by machines. The Semantic Web addresses the first part of this challenge by trying to make the data machine understandable, while Web Mining addresses the second part by automatically or semi-automatically extracting the useful knowledge hidden in these data, and making it available as an aggregation of manageable proportions [1]. Because of the rapid increasing in the amount of stored semantic data and knowledge in various areas, this could be transformed to a perfect target to be mined leading to the introduction of the term "Semantic Web Mining".

In Semantic Web Mining the web pages are mined by the machine can perform better understand the information on the web pages. It is basically mining eXtensible Markup Language (XML) and Resource Description Framework (RDF) documents along with ontologies and metadata to develop an effective Semantic Web. Semantic Web Mining will develop from Web Mining. The goal of Semantic Web Mining is to make easy use of the web. It is also used to provide the web users and machine for better performance of their task.

A general overview of the areas Semantic Web and Web Mining is described in section 2 and 3 respectively. Web mining techniques can be applied to help creating the Semantic Web. The backbone of Semantic Web is Ontology, which is based on RDF and XML. The challenge of Semantic Web is to learn ontologies and instances of their concepts, in an automatic or semi-automatic way. A comprehensive survey on current research and technologies in the area of Semantic Web Mining is presented in section 4. Finally, conclusion is given in section 5.

## 2. Semantic Web

The current World Wide Web (WWW) has a huge amount of data that is often unstructured and only human understandable. Web is rich with information; gathering and making sense of the data in the web is more difficult because the documents of the Web are largely unorganized and unstructured. The nature of most data on the Web is unstructured that only understand by humans, the amount of data is very huge on the web that processed efficiently by machines. If machine can understand the meaning behind this information, it can learn what we are interested in and it help us better find what we want [2]. Therefore, from machine readable data on the Web, to make effectively and efficiently machine understandable is become a challenge. Semantic Web is the solution for this challenge, since it mainly focuses on the data and information.

The Semantic Web was thought up by Tim Berners-Lee, inventor of the WWW, URIs, HTTP, and HTML [3]. The great success of the current WWW leads to a new challenge: a huge amount of data is interpretable by humans only; machine support is limited. Berners-Lee suggests enriching the Web by machine-processable information which supports the user in his tasks. For instance, today's search engines are already quite powerful, but still too often return excessively large or inadequate lists of hits. Machine-processable information can point the search engine to the relevant pages and can thus improve both precision and recall [1].

The Semantic Web initiative from the W3C (World Wide Web Consortium) is a step towards the development of the next generation Web, which is not a new Web, but an augmentation to the existing Web to make it more understandable to machines. The main goal is to express semantic information about Web resources and to store it as metadata along with the resources. This process is known as semantic annotation. The metadata is processed and used by computers to facilitate faster and desired information retrieval.

Data in the Semantic Web is well defined and linked in a way that can be used for more effective discovery,

automation. The goal of the Semantic Web is to develop allowing standards and technologies designed for both user and machines understandable. Semantic web information can support data integration, data discovery, navigation, and automation of tasks. Berners-Lee suggested a layer structure for the Semantic Web [3] as shown in Figure 1.

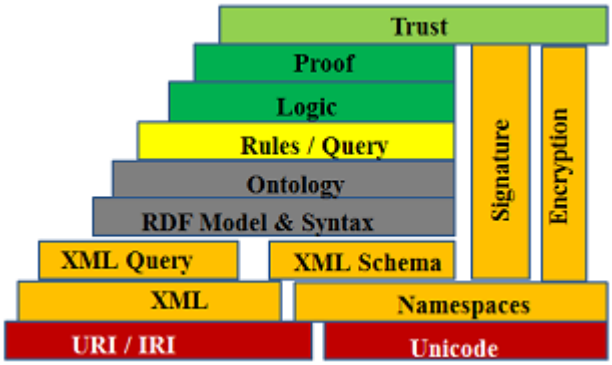


**Figure 1:** Layers of Semantic Web

Many available techniques and models are used to represent and express the semantic of data such as the standard techniques recommended by W3C named XML, RDF, and OWL [4] which are briefly explained below.

**a) XML (Extensible Markup Language)**

XML is the foundation of Semantic Web. It is a language which can store and exchange data in machine readable format between different platforms or devices. XML is only to carry data, not to display data. By enabling users to create their own tags, it allows them to define their content easily. XML Schema is used to describe the structure of the XML document. XML Schema also called as XSD XML Schema Definition. XML Namespace in Semantic Web is used to avoid conflict data or names.

**b) RDF (Resource Description Framework)**

The main constituents of the Semantic Web are metadata and ontologies. The widely used representation structure for metadata is RDF developed by W3C. RDF is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web [6]. The objective of RDF is to represent the metadata of the Web resources by annotating them. Examples of such metadata are the authors, the date creation and the kind of information contained within the document.

RDF defines the data model for the Semantic Web. It has been developed to represent information about Web resources that are uniquely identified via a URI (Unique Resource Identifier). The basic statement is a triple of the form (subject, property, property value) or, equivalently, (subject, predicate, object). The subject of a triple is a resource, which is identified by a URI. The predicate is also denoted by a URI, and the object is either another resource or data type value, also called literal. In case the property value is a resource the property is called an object property, otherwise it is called data type property [7].For example, the statement: "http://www.w3.org/DesignIssues/Semantic.html is created by W3C" is expressed in RDF as a triple as shown in Figure 2.

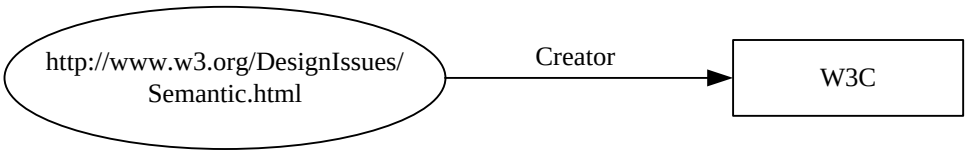


**Figure 2:** Simple Node and Arc Diagram

**c) OWL (Web Ontology Language)**

Ontology is the second important constituent of the Semantic Web. Ontology is a representation of a set of objects, concepts and other entities that are presumed to exist in some domain of discourse and the relationships that hold them. It is application domain dependent and the main purpose of its creation is to share and reuse it as and when required. This requires formal description of the data models in the domain of discourse in terms of a set of concepts, a set of relationships and a hierarchy to relate concepts and relations. Ontology also exhibits reasoning capabilities through the use of axioms, which express universal truths about concepts.

Several ontology languages have been developed during the last few years such as Ontology Exchange Language (XOL), Ontology Markup Language (OML), Ontology Interchange language (OIL) and Web Ontology Language (OWL). These languages facilitate greater machine interpretability of the Web contents.

OWL is based on the top of the RDF and XML based language. RDF is used to represent the rich and complex knowledge about things and their relationship. OWL provides processing information on the web. OWL is a part of web semantics. There are two types of OWL properties i.e. Object properties and Data type properties [2].The OWL is considered a more complex language with better machine-interpretability than RDF. It precisely identifies the resources' nature and their relationships [5]. 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.

A number of tools are also being adopted for the development of the Semantic Web. Examples include Protégé (used for creating ontologies), OntoEdit (used for marking up Web pages with information from external ontologies), LinkFactory (Used for managing ontologies in multilingual terminology) and IBM's Web Ontology Manager (Web-based tool for managing ontologies expressed in Web Ontology Language OWL). A number of

open source tools for creating an integrated system for authoring and searching the Semantic web are also being developed. As the tools for managing ontologies are becoming more prevalent, ontology based Web Mining is also exhibiting tremendous growth. The ontology based Web Mining tends to overcome various problems such as uncertainties associated with the discovered patterns and dynamically changing user profiles etc. by using ontologies to represent the discovered patterns in order to search the right data requested by the users.

**d) SPARQL Query Language**

SPARQL is a standard RDF query language also endorsed by W3C. Given a RDF data graph, SPARQL is an excellent tool to retrieve occurrences of a basic user-specified graph pattern. SPARQL even supports querying conjunctions and disjunctions along with retrieval operators such as projection, distinct, order, limit, and aggregation functions. Thanks to the active development in the Semantic Web community, there are various triple stores, specifically optimized for the storage and retrieval of RDF triples using SPARQL queries.

## 3. Web Mining

Web Mining is data mining techniques for extraction of information from web documents and services. The contents of the web are very dynamic. It is growing at a rapid pace, and the information is continuously updated [9].

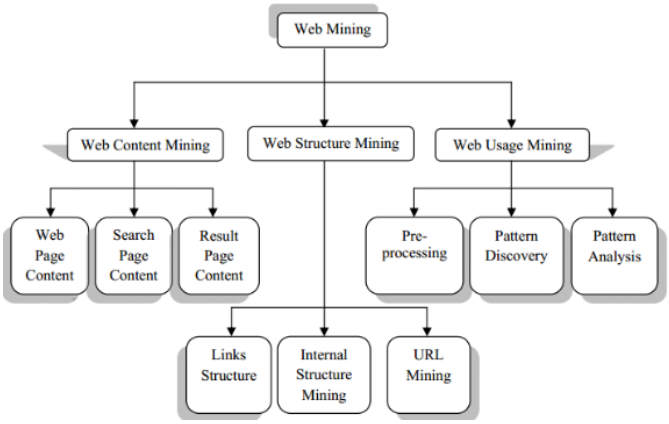


**Figure3**:Taxonomy of Web Mining [19]

Web Mining refers to the discovery of knowledge from Web data that include Web pages, Web links, Web log data and other data generated by the usage of Web data. Web Mining can be broadly classified as: Web Content Mining, Web Structure Mining and Web Usage Mining as shown in Figure 3.Web Content Mining describes the discovery of useful information from the Web contents, data and documents. Web Structure Mining tries to discover the knowledge about link structure connecting Web pages and other Web objects. Web Usage Mining tries to make sense of the data generated by users surfing the net.

Like other data mining applications, Web mining can profit from given structure on data (as in database tables), but it can also be applied to semi-structured or unstructured data like free-form text. This means that Web mining is an invaluable help in the transformation from human-understandable content to machine-understandable semantics [1].

**a) Web Content Mining**

Web Content Mining is the process of extracting information from the contents of Web documents. It examines content of the web pages as well and web searching. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content may be unstructured (plain text), semi-structured (HTML documents), or structured (extracted from databases into dynamic Web pages) [2].

The most successful applications of web content mining are content-based categorization and ranking of web pages, which are adopted by many search engine companies, such as Google, Altavista and Lycos. NLP (Natural Language Processing) researchers are also contributing in developing a new sub field of Web Content Mining called Opinion Mining. It is a recent discipline at the crossroads of Web Mining and Computational Linguistics, concerned not with the topic of a document but with the opinion it expresses [10].

**b) Web Structure Mining**

Web Structure Mining is mostly interested in the hyperlinks of the web pages. Web Structure Mining can be is the process of mining structure information from the Web. It is used to improve the structure of the web pages. Depending upon the hyperlink, the web pages categorize the Web pages and the related information and inter domain level [2]. The area of Web Structure Mining is focusing on the identification of authorities [11]. These are the pages that are considered as important sources of information from many people in the Web community. Another emerging research area related to Web Structure Mining is Link Mining that has its roots at the intersection of various other fields such as Link analysis, Relational Learning, Inductive Logic Programming and Graph Mining [12].

**c) Web Usage Mining**

Web usage mining is the process of extracting information from server logs i.e. user's history and web user behavior. The logs can be examined by client perspective or server perspective. This information takes as input the usage data, i.e. the data exist in in the Web server logs showing the visits of the users to the Web site. Web usage mining is the process of identifying browsing patterns by analyzing the user's navigational behavior [2].

In the Web Usage Mining area, extensive research is being performed towards providing users with dynamic content tailored to their individual interests. This is known as Web Personalization. These Personalization Systems cannot handle heterogeneous objects based solely on their properties without the help of Semantic Web. A key requirement for personalizing the Website is that the server must be able to identify the person visiting that site. Most of the sites that provide Personalization services require a user to enter a login/password combination to open his profile.

## 4. Literacy Survey on Semantic Web Mining

According to [8], Semantics can be utilized for Web Mining in many different ways. For example, hyperlinks in Semantic Web have explicit descriptions and additional information attached with them that can help knowledge engineers in Web Structure Mining. Semantics also facilitate Web Content Mining as the contents of Web pages have latent annotations and metadata attached with them that provide knowledge engineers with more structured inputs required for Web Mining tasks. Web Usage Mining can also get more meaningful patterns from the semantically annotated descriptions of the visited Web pages that can help in improving the Website.

The work presented by [13] gives an overview of where Semantic Web and Web Mining work together, the way how this integration of both areas gives maximum profitable outcomes on WWW and challenges in this area. This paper pointed out the main challenges for Semantic Web Mining are the availability of relevant content, the availability of common ontology, multiplicity of languages, scalability, visualization to reduce information overload, stability of Semantic Web language, ensuring user privacy and understanding the user's natural language queries by Semantic Web.

V. Nebot and R. Berlanga [20] 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" [20], 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 preprocess 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 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.

D. Jeon and W. Kim [21] 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.

The work presented by [22], 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.

Many approaches have been proposed in Semantic Web Mining area that combines Semantic Web data with the data mining and knowledge discovery process. The article presented by [14] gives a comprehensive survey of those approaches in different stages of the knowledge discovery process. As an example, the authors show how Linked Open Data can be used at various stages for building content-based recommender systems. The survey shows that, while there are numerous interesting research works performed, the full potential of the Semantic Web and Linked Open Data for data mining and KDD (Knowledge Discovery in Databases) is still to be unlocked.

There exists a gap between Web mining and the effectiveness of using Web data. The main reason is that we cannot simply utilize and maintain the discovered knowledge using the traditional knowledge-based techniques due to the huge amount of discovered patterns, many noises in discovered patterns and even some useful patterns with uncertainties. S. Suma Singh [19] discussed ontology approaches for building a bridge between Web mining and the effectiveness of using Web data, which tend to automatically construct and maintain ontologies for representations, application and updating of discovered knowledge.

The semantic structure can be built with the pattern or relation results discovered via Web Mining. The work presented in [15] pointed out that the Semantic Web Mining is a recent hot topic in educational research. This paper gives an overview of current applications and techniques of Semantic Web Mining on e-Learning which already became

a base component of education. By applying Semantic Web Mining on educational purposes, especially on distance learning and course management systems where both can be used as a support to traditional education and distance learning intentions. The aim of discovering students' learning model and personalization of services over current e-Learning portals and course management systems are achievable via semantic tools such as Web Services or Semantic Web Agents. Previous applications of Semantic Web Mining on e-Learning systems are explained with their advantages and disadvantages in their study.

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 [23]. 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.

P. Markellou et al. [16] discussed about the application of techniques coming from the new emerging area of Semantic Web Mining in the domain of e-Learning systems and analysed the significant role of ontologies. They expounded and argued about their proposed approach for producing recommendations to users in a given e-Learning corpus. Finally, they concluded with the description of the recommendation engine's operation and presented an algorithm for making effective recommendations. As shown in this paper, the proposed personalization scenario tries to integrate the Semantic Web vision by using ontologies with Web Usage Mining techniques in order to better service the needs and the requirements of learners. The combination of domain ontology and frequent item sets, which include all the information about users' navigational attitude, enhances the whole process and produces better recommendations.

Salah-ddine et al. [17] describe the commonalities of the two areas Semantic Web and Web Mining in order to extract useful and shared knowledge. Their study analyses the merging of trends from both areas including using semantic structures in the Web to enrich the results of Web Mining and to build the Semantic Web by employing the Web Mining techniques. They present that the ontology engineering is very important to solve the problem of the interoperability between web systems by using the ontology learning from web content in order to develop method of extraction knowledge from web data and representation of this knowledge in a machine understandable form.

In the paper presented by [18], MotiurRahman and A. Ferdusee proposed a model to organize the large volume of data over the Web and retrieve the more relevant data to the user. As an implementation of the proposed model, two demo search engines (one for RDF based semantic searching and another for existing searching) are built and two different data sets for testing are used. The comparison results of RDF based searching and keyword based searching are shown in their paper. These results show that the RDF based semantic searching returns more précised data than existing searching, for every set of data. The efficiency of their proposed model is better than the existing searching strategies. In their proposed model, not only traditional web but also RDF based ontology library is considered to organize the data effectively. The lack of comparison result of F-Measure for both semantic and keyword based searching remain as a drawback of their research work.

## 5. Conclusion

In this survey we have studied the two fast developing research areas in Web Mining and Semantic Web. The future of Web Mining will depend on the development of Semantic Web. The combined area of Semantic Web Mining offers new techniques to improve both areas. In this paper a detailed survey of on-going research in Semantic Web Mining and some obstacles faced by researchers has been presented. It is useful to help the researchers who want to contribute something in the area of Semantic Web Mining.

## 6. Acknowledgment

I would express to thank all of my teachers from childhood and nowage. And thanks to Dr. Mie MieThetThwin, Rector, University of Computer Studies, Yangon. I would deeply gratitude to Daw Mu MuMyint, Professor, Head of Department, Faculty of Information Technology Support and Maintenance, for her guidance in my paper. In this case, I must powerful thank all of my friends associated with this paper.

## References

[1] G. Stumme, A. Hotho, B. Berendt, (2006) "Semantic web mining – state of the art and future directions", Journal of Web Semantics, Vol. 4, No. 2, pp124–143. URL: http://www.kde.cs.uni-kassel.de/hotho/pub/2006/JWS2006SemanticWebMining.pdf
[2] J. Sivakumar, K.S. Ravichandran, (2013) "A Review on Semantic-Based Web Mining and its Applications", International Journal of Engineering and Technology (IJET), Vol. 5, No. 1, pp186-192.
[3] T. Berners-Lee, J. Hendler and O. Lassila, (2001) "The Semantic Web", Scientific American, URL: https://www.scientificamerican.com/author/tim-berners-lee-james-hendler-and-ora-lassila/

[4] D. Jeon and W. Kim, (2011) "Development of Semantic Decision Tree", Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications, pp28- 34.
[5] V. Sugumaran and J. A. Gulla, (2012) "Applied Semantic Web Technologies", Taylor & Francis Group.
[6] O. Lassila, Ralph R. Swick, (1999) "Resource Description Framework (RDF) Model and Syntax Specification", W3C. URL: https://www.w3.org/TR/PR-rdf-syntax/
[7] Rettinger et al., (2012) "Mining the Semantic Web", Data Mining Knowledge Discovery, Springer, pp613-662. DOI 10.1007/s10618-012-0253-2
[8] Berendt, A. Hotho, and G. Stumme, (2002) "Towards semantic web mining", Proceedings of First International Semantic Web Conference, pp264–278.
[9] S. Hussain, (2017) "Survey on Current Trends and Techniques of Data Mining Research", London Journal of Research in Computer Science and Technology, Vol. 17, No. 1, pp7-15.
[10] Esuli and F. Sebastiani, (2006) "Determining term subjectivity and term orientation for Opinion Mining", Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, IT.
[11] P. Kolari and A. Joshi, (2004) "Web Mining: Research & Practice", Computing in Science & Engineering, Vol. 6, No. 4, pp49-53.
[12] L. Getoor, (2003) "Link Mining : A New Data Mining Challenge", SIGKDD Explorations, Vol. 5, No. 1, pp85-89.
[13] K. Singh, A. Kumar, A. K. Yadav, (2016) "Semantic Web Mining: Issues and Challenges", International Journal of Engineering Sciences & Research Technology, Vol. 5, No. 8, pp637-644.
[14] P. Ristoski and H. Paulheim, (2016) "Semantic Web in data mining and knowledge discovery: A comprehensive survey", Web Semantics: Science, Services and Agents on the World Wide Web. DOI: http://dx.doi.org/ 10.1016/j.websem.2016.01.001
[15] O. Mustapasa et al., (2010) "Implementation of Semantic Web Mining on E-Learning", Procedia Social and Behavioral Sciences, Vol. 2, pp5280-5283.
[16] P. Markellou et al., (2005) "Using Semantic Web Mining Technologies for Personalized E-Learning Experiences", Proceedings of the Web-based Education, pp1-6.
[17] K. Salah-ddine et al., (2016) "Contribution To Ontologies Building Using the Semantic Web and Web Mining", The International Conference on Engineering & MIS. DOI: 10.1109/ICEMIS.2016.7745329
[18] Md. Motiur Rahman and Ferdusee Akte, (2018) "An Efficient Approach for Web Mining using Semantic Web", International Journal of Education and Management Engineering, Vol. 8, No. 5, pp31-39.
[19] S. Suma Singh, (2018) "A Survey of Web Search from Web Documents Based on Semantic Ontology Technique", American Journal of Engineering Research (AJER), Vol. 7, No. 2, pp284-287.
[20] V. Nebot and R. Berlanga, (2012) "Finding Association Rules in Semantic Web Data," Knowledge-Based Systems, Vol. 25, No. 1, pp51-62. DOI:10.1016/j.knosys.2011.05.009
[21] D. Jeon and W. Kim, (2011) "Development of Semantic Decision Tree," Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications, pp28-34.
[22] H. Liu, (2010) "Towards Semantic Data Mining," Proceedings of the 9th International Semantic Web Conference, Shanghai, pp1-8.
[23] O. Mustapaşa, A. Karahoca, D. Karahoca and H. Uzunboylu, (2011) "Hello World, Web Mining for E-Learning," Procedia Computer Science, Vol. 3, No. 2, pp1381- 1387. DOI:10.1016/j.procs.2011.01.019

## Author Profile

**Khin Mya Nwe,** Lecturer, University of Computer Studies, Yangon