International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 5 Issue: 1 25-Mar-2015,ISSN_NO: 2321-3337

REPRESENTING AND ENRICHING WEB USAGE MINING WITH SEMANTIC INFORMATION A SURVEY

Jagadish kumar.N
Assistant Professor, Velammal institute of technology,Chennai

***ABSTRACT**-- Web mining is the application of data mining techniques to discover patterns from the Web resources. The three main types of web mining are Web content mining, Web structure mining and Web usage mining. Web usage mining is a type of web mining which mainly deals with mining of web usage log files. Web usage Mining aims at discovering insights about the Web resources and their usages from the user's browsing log by applying various data mining techniques such as classification, prediction and clustering. The web logs captured are syntactic in nature with attributes like time of event, URL requested, Status of the Request etc. Adding semantic information such as URL content type, relationship with other URLs is a step ahead in web usage mining. Such a formalization of the Web semantics in web usage mining is gaining more and more importance nowadays. Semantic web describes the web resource with machine process able Meta data. This Meta data can be used to understand the semantics of a particular web page in web usage mining. Semantic web mining is the term used to represent the combination of these two fast growing research areas web usage mining and semantic web. The aim of this paper is to give an overview of current trends in semantic web mining with main focus on web usage mining.*

***Keywords*— Web Mining, Web usage Mining, Semantic Enrichment, Semantic Web mining**

1, INTRODUCTION

Web usage mining explores web access log to find user interest and improvise user experience by implementing recommendations, web page pre-fetching, website adaptation etc. Semantic web treats the World Wide Web as the vast collection of heterogeneous data. Semantic web aims at enriching the web of data with machine process- able information to help the web user get right information with high accuracy and efficiency. As the goal of both of these areas intersects with each other, they can also be used to aid each other [1]. For example the knowledge mined using web content mining can be used to enrich a web resource to provide automatic meta-data enrichment process. The semantic meta-data associated with a web resource (typically a URL) can be used to prune the web usage mining results more meaningfully. The main focus of this paper is to show what role semantic web plays in web usage mining.

The most popular web usage mining technique is to perform OLAP operations by treating the web sites as concept hierarchies and performing roll up drill down on various levels of abstraction. The main difficulty of this approach is conceptualizations often have to be hand-crafted to represent a site that has grown independently of an overall conceptual design, and that the mapping of individual pages to this conceptualization may have to be established [1]. It is thus desirable to take advantage of built in semantic model of a site which not only describes the site but also describes the complex relationship between various sites through ontology representation. Another important web usage mining technique is to mine the frequently accessed web pages or sequentially accessed web pages. This information is then used to provide on line recommendations or site structure improvisation. Usually the mined sequential patterns will be syntactic in nature for example, User who visits page C is likely to visit page D in short duration could be a frequently accessed pattern identified. On a high level this piece of information can be viewed as complete but what if the site C has no relationship with site D and people always mistakenly taken the route of C then opt for site D. Here destination (site D) is

having high importance than the route. This kind of knowledge can be identified easily if the relationship among the links is known prior.

One other web usage mining's central problems is the large number of patterns that are usually found. Among these, how can the *interesting* patterns are identified? For example, an application of association rule analysis to a Web log will typically return many patterns like the observation that 90% of the users who made a purchase in an online shop also visited the homepage a pattern that is trivial because the homepage is the site's main entry point. Statistical measures of pattern quality like support and confidence, and measures of interestingness based on the divergence from prior beliefs are a primarily syntactical approach to this problem. They need to be complemented by an understanding of what a site and its usage patterns are about, i.e. a semantic approach. The workshop on Usage Analysis and the Web of Data (USEWOD2011) was the first workshop in the field to investigate combinations of usage data with semantics and the Web of Data [5].

The section II gives a brief introduction about Web usage mining and its techniques and Section III briefs about Semantic Web constructs section IV elaborates the research trends in integrating the semantic web to enhance the web usage mining.

2, WEB USAGE MINING

In web usage mining the primary resource taken from web is the log record of user's accessed web page URLs most often collected as server logs. Mining for sequential access patterns from user's search log enables interesting relationships between web resources which is normally hidden. For example in an e-commerce site the products displayed may not have any relationship with each other but the user's interest or buying behaviour can create relationships like product B is always preferred by user who buy's product A. Such a market basket analysis can help the business to remodel the site structure or provide combo offers to increase the profit. This type of knowledge is the outcome of sequential pattern mining. In their research work A.C.M. Fong et al [2] proposed a periodic access pattern mining technique. The sequential patterns which are repeated over a period of time are called periodic access patterns. Statistical analysis such as most popular web sites visited, average number of hits for a web page, peak usage level are also outcome of web usage mining. There are several online tools available for doing such kind of analysis.

Clustering groups the web usage transactions and aims to build clusters and categorize users in groups (clusters) who demonstrated similar browsing behavior, also known as user clustering.In Web usage mining, classification techniques used to develop a profile for users who access particular server files based on demographic information available on those

clients, or based on their access patterns. For example classification on WWW access logs may lead to the discovery of relationships such as the following: clients from state or government agencies who visit the site tend to be interested in the page /company/lic.html or 60% of clients, who placed an online order in/company/products /product2, were in the 35-45 age group and lived in Chennai.The navigation patterns discovered can be applied to the following major areas [12].

- Improving the page/site design
- Making additional product or topic recommendations
- Web personalization
- Learning the user or customer behavior
- Web caching or web page pre-fetching by being able to predict the next user access.

Statistical measures of pattern quality like support and confidence, and measures of interestingness based on the divergence from prior beliefs are a primarily syntactical approach to this problem. They need to be complemented by an understanding of what the site is and its usage patterns are about, i.e. a semantic approach [1].

3, SEMANTIC WEB

The Semantic Web is based on a vision of Tim Berners- Lee, the inventor of the WWW. 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-process-able information which supports the user in his tasks [10]. For instance, today's search engines are already quite powerful, but still return too often too large or inadequate lists of hits. Machine-process-able information can point the search engine to the relevant pages and can thus improve both precision and recall. The following steps show the direction where the Semantic Web is heading:

1. Providing a common syntax for machine understandable statements.
2. Establishing common vocabularies.
3. Agreeing on a logical language.
4. Using the language for exchanging proofs.

Berners-Lee suggested a layer structure for the Semantic Web:

i. Unicode/URI
ii. XML/Name Spaces/ XML Schema
iii. RDF/RDF Schema
iv. Ontology vocabulary, Logic Proof,

It follows the understanding that each step alone will already provide added value, so that the Semantic Web can be realized in an incremental fashion. On the first two layers, a common syntax is provided. Uniform resource identifiers (URIs) provide a standard way to refer to entities, while Unicode is a standard for exchanging symbols. The Extensible Mark-up Language (XML) fixes a notation for describing labelled trees, and XML Schema allows

defining grammars for valid XML documents.

The Resource Description Framework (RDF) can be seen as the first layer which is part of the Semantic Web. RDF "is a foundation for processing metadata; it provides interoperability between applications that exchange machine understandable information on the Web." RDF documents consist of three types of entities: resources, properties, and statements. Resources may be web pages, parts or collections of web pages, or any (real-world) objects which are not directly part of the WWW. In RDF, resources are always addressed by URIs. Properties are specific attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource forms an RDF statement. A value is a literal, a resource, or another statement. Statements can thus be considered as object–attribute–value triples. The next layers are the ontology vocabulary and logic. Today the Semantic Web community considers these levels rather as one single level as most ontology allow for logical axioms. Ontology is "an explicit formalization of a shared understanding of a conceptualization". One of the long-term goals of the Semantic Web is to allow agents, software applications and web applications to access and use metadata. A key tool for doing this is a simple protocol and RDF Query Language (SPARQL), which is still in development. SPARQL's purpose is to

extract information from RDF graphs. It can look for data and limit and sort the results. One of the advantages of the RDF structure is that these queries can be very precise and get very accurate results.

4, SEMANTIC WEB MINING

Several studies have shown that Semantic web resources can be used at various stages of web usage mining such as web log pre-processing stage, data transformation stage, data mining stage and finally generating mined rules stage. This survey explores the integration of semantic web information into WUM. The web usage mining's primary resource is user's access logs. The data set includes the browser log, proxy log and Web server log. There are data mining techniques such as classification prediction, proposed to mine these sources individually or collectively.

Depending on web server, web log file data varies on number, type of attributes, and format of log file. W3C maintains standard log file format however custom log file format can also be configured. The important fields are time stamp of the request made, URL requested, the base URL from where this request initiated etc. These are syntactic in nature. In the last decade, several approaches have emerged that use semantics to aid the analysis of Web logs [12] The Semantic Web proposes a way to help computers "read" and use the Web. The recent USEWOD workshop series [5] is a forum for research on the combination of semantics and usage analysis.

Hoxha et al. [6] presented an approach for the semantic formalization of usage logs, which lays the basis for effective techniques of querying expressive usage patterns. In their research each log entry is considered as browsing event (e). Each browsing event is defined with two types of semantic information T_C – Content Type and T_f – type of function that event served.ExampleURLof the event http://dbpedia.org/page/Lyon (Content Type) T_C -{In Proceedings} (Function it serves) T_f – {Informative}. Other examples of content type for a browsing event include person, organization, real estate, education, stock exchange, etc. Whereas, examples of function type could be login, search checkout, reserve, etc. They have used the Log extract from DBPedia and SWDF who's URLs are well defined with domain ontology. They proposed an algorithm that extracted the ontology and RDF descriptions and formed the semantic annotation $\{T_C, T_f\}$ for each URL in the web usage log.

Laura Hollink et al [3] have shown how semantic knowledge can be used to aid several types of analyses of queries on a commercial Web search engine. They showed a way of connecting query log data to the growing universe of public Linked Open Data (LOD). Datasets of the LOD cloud provide valuable background knowledge of the entities occurring in log data. Their method depends on the availability of Linked Open Data on the topics of the queries. Natraj et al [8] proposed a framework that coins the user search log with defined ontology information. They had a client component called onto frame node. Whenever the user searches something in the net using keyword (Ex Apple) this piece of software retrieves semantic information available with the term and save it to the user profile. When the user continues similar tag search (Ex Apple color) then this onto frame node connects with saved user profile and fetch back the result. They have concluded the effectiveness of search results have improved by semantic behavior analysis.

Ramesh et al [9] proposed a variation of A priori algorithm Onto SPM (Onto Sequential Pattern Mining) is defined. The algorithm uses the ontology classification of server log. The pruning step removes the transaction which exceeds the semantic distance threshold. Inference of this study shows promising results in terms of high accuracy and coverage in the online recommendation subject to conditions and assumptions. [7] Yu jiang Liu have done survey on Web Semantics and concluded

Web mining based on semantic networks is used the new semantic to improve the Web mining result and help building the semantic web.

Febian et al [11] In "U-Sem: Semantic Enrichment, User Modelling and Mining Usage Data" on the Social Web investigate the problem of obtaining and mining usage data for Social Web applications. Their work presents a framework for the semantic enrichment and mining of user profiles from usage data obtained from such applications, by incorporating entity extraction, entity identification and topic detection techniques. ACM Fong et al [2] have proposed a system that crates user profile in terms of personalized ontology and have argued that this generated knowledge in terms of semantic meta data automates the usage of it by semantic web applications such as personalized web resources recommendation.

5, CONCLUSION

It is becoming increasingly evident that combining semantic web of data with web mining is future generation of web mining. By introducing semantic information, web usage mining algorithms are performed in terms of ontology individuals instead of web page addresses. This study shows various aspects of these two domains separately as well as combined. The main challenge faced is the availability of a semantic Meta data for all the web resources. To overcome such limitations researchers are trying to combine the content mining and natural language processing techniques to automatically produce the semantic Meta data.

REFERENCES

[1]. Gerd Stumme, Andreas Hotho, Bettina Berendt , "Usage Mining for and on the Semantic Web", ACM digital library. Journal Web Semantics: Science, Services and Agents on the World Wide Web archive Volume 4 Issue 2, June, 2006 .ages 124-143.

[2]. A.C.M. Fong, Baoyao Zhou, Siu C. HuiJie Tang, "Generation of Personalized Ontology Based on Consumer Emotion and Behavior" IEEE transactions on affective computing, VOL. 3, NO. 2, April-June 2012, pp- 152- 164.

[3]. Laura Hollink, Peter Mika,Roi Blanco ,"Web Usage Mining with Semantic Analysis" WWW '13 Proceedings of the 22nd international conference on World Wide Web 2013,pp 561-570.

[4]. http://computer.howstuffworks.com/semantic-web.htm by Tracy V. Wilson site director of How Stuff Works .com with the help of Josh Senecal.

[5]. Usage Analysis and the Web of Data (USEWOD) Workshop Report by Bettina Berendt,Laura Hollink,Vera Hollink 2011.

[6]. Julia Hoxha, Martin Junghans, and SudhirAgarwal ,"Enabling Semantic analysis of user browsing patterns in the web of data", Computing Research Repository abs/1204.2713,2012.

[7]. Yujiang Liu et al, "Study of Semantic Web Usage Mining", International Conference on Communication Systems and Network Technologies 2013.

[8]. Nataraj J, Mural Krishnan P, Praveen Godfrey, Arun S.N. "Extraction of Ontological Information based on Semantic Analysis" 2013, pp 3340-3344.

[9]. C.Ramesh, Dr. K. V. Chalapati Rao, Dr.A.Goverdhan "A Semantically Enriched Web Usage Based Recommendation Model", Transactional International Journal of Computer Science & Information Technology (IJCSIT) International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 5, Oct 2011.

[10]. Tim Berners-Lee. Semantie Web Arehitechture. http://www.org/2000/Talks/1206-xmlZk-tbl/slide10-0.html

[11]. Fabian Abel, Ilknur Celik, Claudia Hauff, Laura Hollink, Geert-Jan Houben, "U-Sem: Semantic Enrichment, User Modelling and Mining of Usage Data on the Social Web", arXiv:1104.0126v1 [cs IR] 1st Apr 2011.

[12]. B.Lalithadevi* A.Merry Ida W.Ancy Breen, "A New Approach for Improving World Wide Web Techniques in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering 3(1), January – 2013

BIOGRAPHY:

Am Jagadish kumar.N currently working as an assistant professor in Velammal institute of technology .Previously am a corporate trainer in a leading MNC. My area of Interest is Knowledge and data engineering.

.