Web Log Mining Based on Consumer Behaviour Analysis using Jena

Renjini Ram
Computer science and Engineering Department
Government Engineering College Idukki
Kerala,India

Philumon Joseph
Assistant Professor ,CSED
Government Engineering College Idukki
Kerala,India

***Abstract*—Web log mining is on its booming stage as far as the time is concerned. New definitions of data storage has been emerged as BIG DATA. All these years the weblog mining has been done for market analysis, data analysis ,financial management etc… . Another perspective on consumer behavior analysis through web mining is done here with Jena. Jena is a kind of distributed file system where all data have no index like normal relational kind of storage other than mapping . Consumer behavior analysis actually a kind of market basket analysis and is done to fire up the impulsive buying of a customer through recommending items which are probably the customer wants. Impulsive buying on recommendation is surprisingly higher and there lies the point.**

Keywords—semantic web; weblog minig; Association rule minig; Jena;

I. INTRODUCTION

People on these days relays internet for their daily needs and all these contents are available at one finger tip touch. From their data accessing details we can found that most of the time they follows a particular pattern . The data analyst can exploit these kind of data for content recommendation . Over these years so many content recommendation techniques were evolved among these the one using the association rule mining has got the credibility. Especially fuzzy association rule mining combined with case based reasoning[1] is done by Wong et.al [2]. From this we can predict futuristic we access patterns.

The conventional web system is upgraded from web 2.0 to web 3.0. Web3.0 so called Semantic web is mainly focusing on this content recommendation. Its futuristic scope is unpredictably productive. To discover the information for this content recommendation we need the personalized and globalised details of the content or likely be a content which is going to be recommended . All these details constitute a pattern and which is tend to show the periodicity. Tendency to repeat this patterns. Studies shows that people tend to search for a particular thing when they are recommended other than searching by self. To capture these consumers periodical access patterns web usage mining is performed. There are many web usage mining techniques like Association rule mining , Support vector based mining and Regression based mining . From the support and the confidence value a content lattice is created . In this paper the web usage mining is performed using association rule mining. In this some fuzzy variables are set and from that formal concept lattices are created.

II. RELATED WORKS

A. *Semantic web*

Semantic Web is an extension of the current World Wide Web in which information is given in a well-defined meaning a.k.a web 3.0According to [1] Semantic web is an effort to enhance current web so that computers can process the information presented on WWW, **interpret and connect it, to help humans to find required knowledge.** Semantic Web translates the given unstructured data into knowledgeable representation data thus enabling computers and people to work in cooperation. Semantic Web is information in machine understandable form. It is also called as Global Information Mesh (GIM). The Semantic Web is distributed and heterogeneous, has brought the evolution of the Web to a higher level. It is meant to improve its usability as a medium for collaboration and the second to ensure that its contents can be understood by machines. Providing annotation data will help this second aim. Numerous tools and applications for Semantic Web technologies have recently become available.

Universal Resource Identifier(URIs) are used to identify and locate resources or anything in the web . The URI, which is considered to be the foundation of the Semantic Web, is used to give a unique name to each resources and relation between the resources .URI is different from URLs. URLs never indicate relations but locate resources .A URN is a URI that identifies a resource by name in a particular namespace. **uniform resource identifier** (URI) is either a uniform resource locator (URL), or a uniform resource name (URN), or both. Thus we can infer it as No all URIs are URLs, but all URLs are URIs .Unicode is the standard for computer character representation which is commonly used. A URL is a URI that, in addition to identifying a web resource .

B. *Ontology*

Ontology is the term used to refer to the shared understanding of some domain of interest which may be used as a unifying framework to solve the above problems in the above described manner . An ontology necessarily shows some sort of world view with respect to a given domain .The world view is often conceived as a set of concepts e.g. entities attributes processes their definitions and their inter relationships this is referred to as a conceptualisation. An ontology[5],[6],[7] contains some sort of world view with respect to a given domain

explicit ontology may take a variety of forms but necessarily it will include a vocabulary of terms and some specification of their meaning ie definitions . The degree of formality by which a vocabulary is created and meaning is specified varies considerably. Four somewhat arbitrary points along what might be thought of as a continuum are highly informal expressed loosely in natural language.*Semi informal* expressed in a restricted and structured form of natural language greatly increasing clarity by reducing ambiguity eg the text version of the Enterprise Ontology .*Rigorously formal* meticulously defined terms with formal semantics theorems and proofs of such properties as soundness and completeness

Ontologies can be broadly divided into two main types: lightweight and heavyweight. Lightweight Ontologies involve taxonomy (or class hierarchy) that contains classes, subclasses, attributes and values. Heavy weight Ontologies model domains in a deeper way and include axioms and constraints .

C. RDF

RDF [2] is a basic data model for writing simple statements about Web objects (resources and edges) specified with URIs.

RDF Model has three components: Resource, Property and Statement. Resources and Edges(that connects resources) are URIs . Both XML and RDF follow same syntax in writing properties. RDF is a data model can be represented as a graph. It uses URIs to represent each item.(some Uniform resource locator).In RDF graph there is resources ,edges , literals and blank nodes. Blank nodes are resources with out URI. RDF is a set of all URIs for the edges that make up the graph . Edges relates thing and gives a meaning ie; specific URI specific language. The RDF graph contains

statement or a triple : It is a 3-tuple of the form (Subject,Predicate,Object)Subject and Predicate are URIs and Object is URI or Literal.Ie; RDF is a collection of triples **RDF named graph or quards :** It is a collection of RDF statements that has a name. It is a tuple of the form (named graph,Subject,Predicate,Object). **RDF Schema :** It is a collection of classes with certain properties using the RDF extensible knowledge representation language, providing fundamental things for the description of ontologies, a.k.a called RDF vocabularies, intended to structure RDF resources. These are stored as triples and saved in a triplestore and we can reach them with the query language SPARQL (which is a modelling language), it also provides a reasoning framework for inferring types of resources. **RDF vocabulary :** It is a set of URIs of the edges that constitutes RDF graph. Edges relates things in graph and give it meaning. Using a specific URI means using specific languages . In order to share data between the semantic web must agree on a common vocabulary .

D. Fuzzy powered association rule mining

When the accessed web usage contents are disjoint having no relation between them we need create some co relations by examining some fuzzy relationships between them . Fuzzy variables are not like binary variables having zero or one discrete values rather than varies between 0 and 1. It always gives a fractional value between zero and one showing that how much the variable is related to the truth value and the false value. Ie; in the traditional methods search for a value which is true or false not the relativity showing how the value is closer to its supermom or vice versa. For the prediction accuracy, the fuzzy set association rule mining shows a better
result.

E. JENA

To programming with RDF and OWL in java platform we can use JENA[7] which is an open source project initiated by HP and can be downloaded from http://jena.sourceforge.net/ . There are other frameworks available for programming with RDF and OWL is Sesame and OWL API. For .NET platform SemWeb , for PHP platform RAP and for C the RedLand is used to operate on RDF and OWL. JENA is a java frame work for building Semantic web applications. JENA includes rule based inference engine and it is open source . The JENA framework includes an RDF API, reading and writing RDF in RDF/XML,N3,N-Triples,Turtles,An OWL API, In memory and Persistent storage, SPARQL and RDQL query languages for RDF.

III. WEB LOG MINING

Web logs are semantically enriched URLs and it is collected from a traditional web server and it is preprocessed in order to avoid that URLs having unsuccessful data requests. After this data cleansing process the next is to find all the users and the sessions.All these process are done to avoid unnecessary data and to identify personal access sessions for each users. In this paper the web log mining is done for the purpose of content recommendation. Content recommendations are of two types the content- based recommendation retrieves other documents similar to those the user liked earlier. Collaborative recommendation retrieves documents liked by other people similar to the user. In the filter based recommendation system the filtering for the recommendations are done using User – based filtering system or by the Item based filtering.

In this paper two sections of operations are performed like first a personalized user oriented search pattern have been created and also a global pattern is also created. On theses Ontologies the SparQL queries are passed to get the apt recommendation item and the user. All the Ontologies on personalized user weblog are created using Apache jena and the Ontology part of the web mining is done in the Jena System.

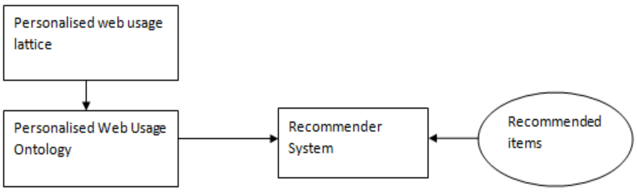
Personalised web usage lattice
Personalised Web Usage Ontology
Recommender System
Recommended items

Figure 2. Web log based content recommendation System

A. Personalised Web Usage Ontology

This ontology is created from the weblog dataset. By setting the fuzzy variables for the resource set as sports , social network sites and games and another property set considering morning ,noon and evening we can calculate the support and confidence values for each item according to the user. This values then further results in lattice discrete structure .
Consider a web usage tuple (G, P,R) where G = All the web usage logs, P and R are the property and resource set. Consider H⊆ P∪ R and if B is not empty then

Sup(B) = $(up(g) * ur(g)/|G|$ for all g in B.

Hence the fuzzy support of the web usage context is calculated After that the fuzzy confidence of the that support is calculated like

Conf(B) = Prob((B∩ P)|(B∩R)) = Sup(B)/Sup(B∩ P)

Confidence value closer to one or larger than is taken into consideration and then the concept lattice is created

IV. IMPLEMENTATION

In this paper proposes a method to implement semantic web through creating ontology. Semantic web is a collection of ontology or RDF[8..10] schemas . In semantic web the information in the web is processed according to their meaning and retrieves the result. The get the domain knowledge .

Here a collection of URL has been selected according to the criteria having sports , games and facebook.Each got a fuzzy variable as morning noon and evening . According to the URl we separates it periodic attributes and resource attributes from these attributes we can calculate the fuzzy values.

From these fuzzy values the support and confidence values of the resources are obtained . According to the association rule criteria the resource combinations having the high confidence value should be the most recommending item .

To develop ontology the following steps can be adopted. Semantic web is a collection of different knowledge base and these ontologies are perfectly made by exploring goal and scope of the ontology in the particular domain. Then define ontology domain conceptual model also identify the classes , relations and attributes for the ontology. In the next step instance creation of ontology is performed and the ontology consistency verification is done by using a reasoner. After completing the competency question validation through SPARQL the final ontology were ready. All these operations can be done by using Jena.

In jena there are statement centric methods for manipulating an RDF model considering all as a set of RDF triples.

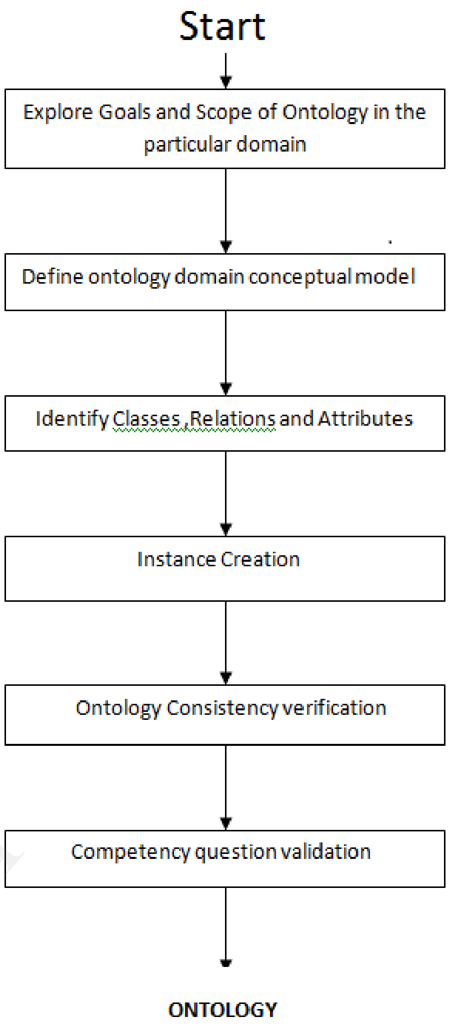
Start
Explore Goals and Scope of Ontology in the particular domain
Define ontology domain conceptual model
Identify Classes ,Relations and Attributes
Instance Creation
Ontology Consistency verification
Competency question validation
ONTOLOGY

Figure 3 : Ontology creation Block Diagram

Also jena constitutes of resource centric and cascading method calls for more convenient programming. Bag, Alt and Seq are examples of RDF containers and all these are supported by jena.Integrated parsers and enhanced resources are some other features of jena. It also supports typed literals. For creating Graphs and Statements in jena the following statements can be used.In jena an RDF graph is called Model sample code will look like :

Model model = ModelFactory.createDefaultModel(); To create a resource , property and also to add that property to that resource we can use the following statements .

Resource renjiniRam = model.createResource("http://
somewhere /RenjiniRam");

Property hasName = model.createProperty("http://
example.com/terms#hasName");

Add the property hasName to the resource renjiniRam

renjiniRam.addProperty(hasName, "Renjini Ram");

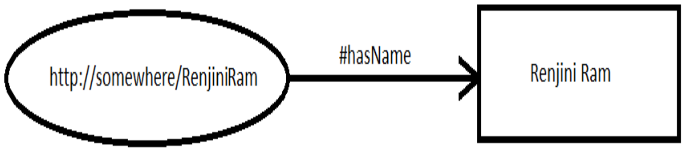

Figure 4 : Sample RDF Graph

To perform read and write RDF using InputStream and OutputStream the following syntax can be used read(InputStream os, String base, String RDFSyntax) and write(OutputStream os, String base, String RDFSyntax) . For navigating and manipulating through RDF the following code can be used. model.getResource(uri), stat.changeObject (some_ value_or_URI), stat.getObject(). The validation using the competency question needs an ontology and needs ontology instances. On this Ontology the SPARQL queries are passed as the competency question. If the results are logically right then it is considered as the basement ontology for the semantic web.

V. RESULT AND ANALYSIS

This project results in a RDF document which further can be used as the RDF schema .The semantic searches can be done on this RDF. Here a new base ontology is created and new nodes can be further attached to it. Querying are done with SPARQL and also the reasoning is also performed. Examples for the semantic web are Hakia and Sindise and they operates on RDF. Its obvious that we may interest to find out similar soursce present in WWW .This semantic web provides semantic extensions to find similar data content and not just by arbitrary descriptions.

For analysis there is a java semantic measures library and it is open source.It is generic and can be used for multiple ontologies. The analysis is performed here is the semantic similarity measures ie: the likeliness of terma words and documents. The likeliness of compared object is based on their meaning or semantic content. This semantic similarity can be estimated for instance by defining a topological similarity. By using ontologies to define a distance between terma and concepts. Topological similarity are of two types Edge based and Node based.

VI. APPLICATION

The Ontology and semantic web have many applications in NLP . In the area emotion recognition and AI these are also used [11..14]. Content recommendation is also an application of semantic web. Web usage mining is also a vast area of application for the semantic web [15..20].

VII. CONCLUSION

In the modern world everybody relay on internet for each and every application . Technology savvy people search everything on internet and the fluid web or the semantic web is a solution for that. Since the weblogs have been distributed between server clusters and usage helps us to easily implement the weblog mining system. Hadoop is a convenient option if we Need to process Multi Petabyte Datasets and it is expensive to build reliability in each application. Weblog mining on these days has been got so much scope and will help the further contributions on content recommendation

VIII. ACKNOWLEDGMENT

This survey is conducted as part of Renjini Ram's Mini Project. Comments and suggestions from Prof. Unnikrishnan and Shabarish are gratefully acknowledged.

V. REFERENCES

1. L.A. Zadeh, "Fuzzy Logic and Approximate Reasoning,"Synthese, vol. 30, pp. 407-428, 1975.
2. C. Wong, S. Shiu, and S. Pal, "Mining Fuzzy Association Rules for Web Access Case Adaptation,"Proc. Fourth Int'l Conf. Case-Based Reasoning, Case-Based Reasoning Research and Development,pp. 213-220, 2001
3. Aurona Gerber, Alta van der Merwe, and Andries Barnard "A Functional Semantic Web Architecture". *Meraka Institute and University of South Africa (Unisa),Pretoria, South Africa 2006.*
4. Mike Uschold , Michael Gruninger, "Ontologies: Principles Methods and Applications". *Volume 11 Number 2 June 1996. (references).*
5. Alexander Maedche and Steffen Staab , "Ontology Learning for the Semantic Web".
6. Michael Kifer1, Jos de Bruijn2, Harold Boley3, and Dieter Fensel2 , "A Realistic Architecture for the Semantic Web" , 2006
7. Amit Sheth, and Cartic Ramakrishnan Semagix and LSDIS lab, University of Georgia , "Semantic (Web) Technology In Action: Ontology Driven Information Systems for Search, Integration and Analysis" , Slightly abridged version appears in *IEEE Data Engineering Bulletin, Special issue on Making the Semantic Web Real, U. Dayal, H. Kuno, and K. Wilkinson, Eds.* December 2003.
8. Sergej Sizov, University of Koblenz-Landau, "What Makes You Think That? The Semantic Web's Proof Layer" , *university of alberta. october 1, 2009*
9. Matthias Hert , " Semantic Web Engineering",*University of Zurich 2010.*
10. Hans-Jörg Happel*) and Stefan Seedorf†), " Applications of Ontologies in Software Engineering ", 2007.
11. Dieter Fensel and Frank van Harmelen, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, "OIL: An Ontology Infrastructure for the Semantic Web " , IEEE Intelligent Systems , March/April 2001.
12. Roberto Poli • Michael Healy • Achilles Kameas , "Theory and Applications of Ontology: Computer Applications" ,Springer 2010
    [1]
13. Andrew U. Frank , "Ontology: a consumer's point of view ".1996.
14. Miguel Angel Salichs, Member, IEEE, and Marı´a MalfazA ,"New Approach to Modeling Emotions and Their Use on a Decision-Making System for Artificial Agents", IEEE transactions on affective computing, vol. 3, no. 1, january-march 2012.
15. Alexandra Balahur, Jesu´s M. Hermida, and Andre´s Montoyo, "Building and Exploiting EmotiNet, a Knowledge Base for Emotion Detection Based on the Appraisal Theory Model" ,IEEE transactions on affective computing, vol. 3, no. 1, january-march 2012.
16. A.C.M. Fong, Senior Member, IEEE, Baoyao Zhou, Siu C. Hui, Jie Tang, Member, IEEE, and Guan Y. Hong, Member, IEEE ," Generation of Personalized Ontology Based on Consumer Emotion and Behavior Analysis " , IEEE

transactions on affective computing, vol. 3, no. 2, april-june 2012.

17. Baoyao Zhou, Siu Cheung Hui, Alvis. C. M. Fong, "A Web Usage Lattice Based Mining Approach for Intelligent Web Personalization" , J. Web. Infor. Syst. 1 (1), March 2005. Troubador Publishing Ltd.
18. Gaihua Fu, Christopher B. Jones and Alia I. Abdelmoty ,"Building a Geographical Ontology for Intelligent Spatial Search on the Web" , unpublished.
19. Jeff Z. Pan," A Flexible Ontology Reasoning Architecture for the Semantic Web ", IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No. 2, February 2007.
20. Mohammad Mustafa Taye , "Understanding Semantic Web and Ontologies: Theory and Applications" , Journal of Computing, Volume 2, Issue 6, June 2010, ISSN 2151-9617.
21. Benjamin Heitmann, Sheila Kinsella, Conor Hayes, and Stefan Decker , "Web Usage Mining for Semantic Web Personalization" ,2009
22. Christopher B. Jones , "Spatial Information Retrieval and Geographical Ontologies An Overview of the SPIRIT Project" , unpublished.