WEBMINING AND SEMANTICS WEBMINING:REVIEW

[1] Anita Tripathi

(Assist.Prof. AIHE, ,Kanpur)

[2]Ms.Prachi Kushwaha

(Assist.Prof. AIHE, Kanpur)

[3] Mr.Ashwani Shakya

(Assist.Prof. AIHE, Kanpur)

Abstract—This paper describes about Semantic Web Mining. The Purpose of this paper is to focus on how semantic web technologies can be used to mine the web , for relevant information extraction. Semantic Web Mining is about combining the two emerging research areas Semantic Web and Web Mining. Researchers work on improving the result off web mining by using semantic structure in the web and make use of Web Mining techniques for building the Semantic Web. In this manner both technologies are playing vital role to each other. Semantic Web adds structure to the meaningful content of Web Pages ; hence information is given a well defined meaning; which is both human readable as well as machine-processable. This paper gives an overview of where the two areas meet today, and sketches ways of how a closer integration could be profitable.

Keywords- Semantic Web; Web Mining; Ontologies; Knowledge engineering; World Wide Web

I. INTRODUCTION

On the Web most data are 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. Due to huge data it is very difficult to access relevant information when required. To overcome with this use of two fast- developing research area Semantic Web and Web Mining both are fruitful in this aspects. Semantic Web and Web Mining build on the success of the World Wide Web(WWW).Semantic Web is useful to make the data machine-understandable, while Web Mining extract the useful knowledge hidden in these data, and making it available as an aggregation of manageable proportions.

Web Mining techniques can be applied to help create the Semantic Web. A backbone of the Semantic Web are onto logies.The challenge is to learn ontologies and its concepts. In Semantic Web ontologies can be used to improve the process and result of Web Mining. Recent developments include the mining of sites that become more and more Semantic Web sites and the development of mining techniques that can use the power of Semantic Web knowledge representation .The tight integration of Semantic Web and Web Mining will greatly increase the understandability of the Web for machines, and will thus become the basis for further generations of the intelligent Web tools. In this manner both emerging technologies are complimenting each other to fetch out relevant information from large pool of data using various technologies. Type Style and Fonts Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word Times. Avoid using bit-mapped fonts if possible. True-Type 1orOpenTypefontsarepreferred.Pleaseembedsymbol fonts, as well, for math, etc. processor, please use the font closest in appearance to

II. SEMANTICWEB

SemanticWeb is an extension of current web, in which information is given well-defined meaning, better enabling computer and people to work in cooperation [1] .

The Semantic Web is based on a vision of Tim Bemers- 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. Bermers-Lee suggests to enrich the Webby machine-process able 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 process able information can point the search engine to there levant pages and can thus improve both precision and recall. For instance, today it is almost impossible to retrieve information with a keyword search when the information spread over several pages.

Semantic web aims at offering a solution, capturing and exploiting the meaning of terms to transform current web from information-presentation platform to a platform that focuses on understanding and reasoning with the information [2].

The following steps show the direction where the Semantic Web is heading:
1. Providing a common syntax form a chine understandable statements.
2. Establishing common vocabularies.
3. Agreeing onalogical language.
4. Using the language for exchanging proofs.

In Semantic Web Ontology approach is needed to describe. Anontology is"an explicit formalization of a shared understanding of a conceptualization"[3].

III. WEBMINING

The development of World Wide Web and its usage grows, it will continue to generate evermore content, structure, and usage data and the value of Web mining will keep increasing. Research needs to be done in developing the right set of Web metrics, and their measurement procedures, extracting process models from usage data, understanding how different parts of the process model impact various Web metrics of interest ,how the process models change in response to various changes that are made-changing stimuli to the user, developing Web mining techniques to improve various other aspects of Web services, techniques to recognize known frauds and intrusion detection.

Web Mining is the application of data mining techniques to the content, structure, and usage of Web resources. It is thus"thenontrivialprocessofidentifyingvalid,previouslyunknown,andpotentiallyusefulpatterns"[4]in the huge amount of these Webdata, patterns that describe them in concise form and manageable orders of magnitude. 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.

Three areas of Web Mining are commonly distinguished: Content mining, structure mining, and usage mining[5,6,7]. In all three areas, a wide range of general data mining techniques, in particular   association rule discovery, clustering, classification, and sequence mining, are employed and developed further to reflect the specific structures of Web resources and the specific questions posed in Web Mining.

SEMANTIC WEB MINING

The process of mining for semantic data involves several processes namely, crawling the web for semantic web documents, extracting and analyzing information from this data, clustering the semantic data for later retrieval purposes, and its scope for the future which involves.

Enhancing the capability of information extraction systems by adding reporting functionality which involves tracking changes in information over time.

The task of mining the Web for Semantic Data essentially consists of crawling the web and finding Semantic Web Documents, which are stored in the form of RDF, OWL, FOAF ,RSS, etc at various locations. This leads us the idea of designing a robust RDF crawler. *Crawling the* semantic web is essentially identical to crawling the HTML content web-it's simply a case of choosing one or more starting points, downloading a resource and following the pointers in it to further resources[8].The difference between gathering HTML and RDF data is that RDF has a well defined mechanism for merging multiple RDF models. We many combine any number of RDF models to produce a single unified model. Hence instead of performing the task of building a database of keywords and links to locations where HTML representations related to those keywords can be found, the RDF crawler can create combined model for all the semantic data found. The major advantage of this union of models is that the model now becomes a rich resource of information. That is one document contains the combined information of the all the separate documents which contain fragments of data. Some of the design

considerationswhileimplementingsucha crawler couldbeResourcePoolingtoavoidoverload ontheserver, Gathering URLsfromcertaintargetsinRDFrepresentationE.Gthe

<rdf:seealso>   triples that contain additional information about a document and mapping of the ontologytothe data.After downloadofSemantic Datais complete, we nowhave tomove tothesecond partofthe process that is extraction andanalysis of information from the data, one of the most efficient ways to extract information fromRDF graphsarebyusingRDQL(RDFdataquerylanguage).   RDQLisnowsupportedby   manypopularRDFAPI frameworkssuchasJena1.Figure1showstheproposed desig.
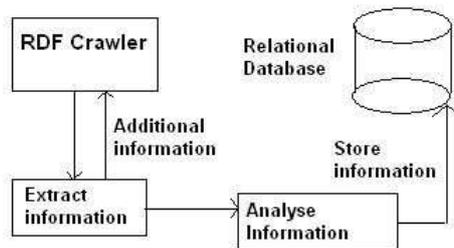
V. RELATED WORK

Staab[10]presentsworkontheOntotextRDFcrawler which downloads interconnected fragments of fragments of RDF from the Internet and builds a knowledge base from its data. A host of URIs to be retrieved as well as URI filtering conditions are maintained at every phase of RDF crawling. This is done in order to download the resources containing RDF iteratively. Toenable embeddingin othertools, RDF Crawler provides a high-level programmable interface (Java API). Other work done in this field includes the Hackdiary

2RDF Crawler which is a multithreaded java implementation capable of downloading simultaneously from many sources whiletheaggregation threaddoes theprocessing. Itbuilds amodel thatremembers theprovenanceoftheRDF and takes caretodeleteandreplacetriplesifithitsthesameURL twice.Hence the data is up-to-date all the times even after many runs.

VI. FUTURE WORK

The InformationExtractionsystemsE.G Armadillo workontheprincipleofutilizingtheredundantinformation onthe webby usingmultiplecitationsasawayofvalidatingthe data. The valid data is thenused to bootstrap the annotation processbyusingIEAnnotationenginessuchasAmilcare [11].Henceproducingmachine- readablecontentforthe SemanticWebi.e.SemanticWebDocuments.Armadillo outputsRDF documents aftercrawling theweb. Armadillo iscurrentlyabletolearnovertheHTMLcontentofthe World Wide Web. However if the IE system is able to learn from semantic data E.G RSS and XML feeds which are now increasingly beingprovidedbymostwebsitesandimplementamechanismtotrackthechangesinformation over this data.

VII. CHALLENGES

Extracting anontology fromthe Web is achallenging task.Onewayistoengineertheontologybyhand,but thisis expensive.In[14],theexpression*ontologylearning*was coined for the semi-automatic extraction of semantics from theWeb.There,machinelearningtechniqueswereusedto improve the ontology engineering process andto reduce the effortfortheknowledgeengineer.Ontologylearning.

exploitsmanyexistingresourcesincludingtexts,thesauri, dictionaries,anddatabases(see[15]asanexample ofthe use ofWordNet). It builds on techniques from Web content mining,anditcombinesmachinelearning techniqueswith methodsfromfieldslikeinformationretrieval[16]and agents [17,18], applying themto discover the „semantics" in the data and to make them explicit. The techniques produce intermediateresultswhichmust finallybeintegratedina machine understandable format, e.g., ontology.

VIII. CONCLUSION

Inthispaper,wehavestudiedthecombinationofthetwo fastdevelopingresearchareasSemanticWeband WebMining.SemanticWebMiningaimsatcombiningthetwofast-developingresearchareasSemanticWeb andWeb Mining.Thissurveyanalyzestheconvergenceoftrends fromboth areas: More and more researchers are working on improving theresults of Web Mining byexploiting semantic StructuresintheWeb,andtheymake useofWebMining techniques for building the Semantic Web.). Semantic Web isusefultomakethedata machine-understandable,while WebMiningextracttheusefulknowledgehiddeninthesedata,andmaking itavailableasanaggregationof

manageableproportions.Discussionoftechnologyusedin Semantic Web Mining which includes RDF crawler.Using SemanticWebMiningrelevantinformationfromvarious huge pools of data can be fetch. There are Various research work going on Semantic Web Mining which is discussed.

REFERENCES

[1]   Schwartz,DG,2003.Fromopenissemanticstothesemanticweb:TheRoadAhead,IEEE Intelligent systems, pp 52-58.

[2]   McGuinness,DL,Fikes,R,Hender,J,Stein,LA,2002. DAML+OIL:
Anontologylanguageforthesemanticweb,IEEEIntelligentsystems, pp72-80.

[3]   T.R.Gruber,TowardsPrinciplesfortheDesignofOntologiesUsedforKnowledgeSharing,in:N. Guarino, R. Poli (Eds.), Formal Ontology in Conceptual AnalysisandKnowledge   Representation,   Kluwer, Deventer, Netherlands, 1993.

[4]   U.M.Fayyad,G.Piatetsky-Shapiro,P.Smyth,Fromdataminingtoknowledgediscovery,in:U.M.Fayyad,
G. Piatetsky-Shapiro, P.Smyth, R.Uthurusamy(Eds.),AdvancesinKnowledgeDiscoveryandDataMining, AAAI/MIT Press, Cambridge, MA, 1996, pp. 1–34.

[5]   O.R.Za¨ıane,Fromresourcediscoverytoknowledge discovery ontheinternet.TechnicalReportTR1998–13,SimonFraserUniversity1998.

[6]   R.Kosala,H.Blockeel,Webminingresearch:A survey, SIGKDD Explorations 2 (1) (2000).

[7]   J.Srivastava,R.Cooley,M.Deshpande,P.-N.Tan,   Webusagemining:Discoveryandapplication   ofusage patternsfromWebdata,SIGKDDExplorations1(2) (2000) 12–23.

[8]   Biddulph(2003)CrawlingtheSemanticWeb.BBCLondon, UnitedKingdom

[9]   Zamir,O.Etzioni.WebDocumentClustering:AFeasibilityDemonstration,Departmentof Computer Science and Engineering,UniversityofWashington.

[10]   Staab,K.Apsitis,S.Handschuh,H.Oppermann,(2004)SpecificationofanRDFCrawler.

[11]   Ciravegna,S.Chapman,A.Dingli,andY.Wilks.
LearningtoHarvest Information for theSemanticWeb. DepartmentofComputerScience,TheUniversityOf Sheffield.

[12]   K.   Elissa,   "Title   of   paper   if   known,"   unpublished.Baader,F,Calvanese,D,Giacomo,GD, Fillottrani,P,Franconi,E,Grau,BC,Horrocks,I,   Kaplunova,A,Lembo,D,Lenzerini,M,Lutz,C,   Moller,R,Parsia,B,Patel-Schneider,P,Rosati,R, Suntisrivaraporn,B,Tessaris,S,2006.Formalismsfor Representing Ontologies: State of theArtSurvey,TONES.

[13]   B.LeGrand,M.Soto,Xmltopicmapsandsemanticwebmining,in:G.Stumme,A.Hotho,B. Berendt (Eds.), Semantic Web Mining,Freiburg, 3rd September,
2001,   12th   Europ.   Conf.   on   MachineLearning(ECML"01)/5thEurop.Conf.onPrinciplesand Practice of Knowledge.

[14]   A.Maedche,S.Staab,OntologylearningforthesemanticWeb, IEEE Intelligent Syst. 16 (2) (2001) 72–79.

[15]   A.Maedche,S.Staab,OntologylearningforthesemanticWeb, IEEE Intelligent Syst. 16 (2) (2001) 72–79.

[16]   G.Paaß,   J. Kindermann, E. Leopold, Learningprototypeontologiesbyhierarchicallatentsemantic analysis. In [166], 2004, pp. 49–60.
A.Maedche,OntologyLearningfortheSemanticWeb,Kluwer,2002.