RESEARCH ON SEMANTIC WEB MINING: A REVIEW

[1]R.B. Ghayalkar and [2]S.R. Sirsat

[1]Department of Computer Science & I. T. Shri R. L.T. College of Science, Akola, MS, India.

[2] Head Department of Computer Science, Jijamata Mahavidyalaya Buldana, MS

rambg29@gmail.com, sandeep_sirsat@rediffmail.com

ABSTRACT

World Wide Web is become largest digital database in the world but this information is generally understandable to human not by machines. Currently information retrieval is based on key-words so sometime it unable to retrieve the relevant information according to the users query.

In this paper presents the review of semantic web mining ie three research areas semantic web, web mining & multiple agent system. Here focuses on semantic to understand and improve the web services and improving the relevancy in retrieved documents.

***Keywords**: Web mining, Semantic Web, (SW), Multi-agent System, Ontology, Information Retrieval(IR), RDF*

Introduction

Information on World Wide Web is exponentially increasing and this process is continuous growing on. Nowadays billions of web pages and their sources which contain hugs amount of information which leads to "Information Overload". Due to this overloading of information it is very difficult for search engines for proper indexing the web pages and retrieving exact relevant information. It arise the problem of ranking irrelevant documents as relevant. Sometimes web users are not able to express their ideas or requirement and don't know that how to get the accurate information.

These problems could be overcome by some extension of information retrieval techniques, filtering approaches and web text mining.

Moreover, the existing search engines are keyword-based. The Web pages matching the given keyword are highly ranked which paved way for Web spamming. Web spamming is the act of inducing malicious bias to search engines so that the web pages are ranked much higher than they deserve. This leads to poor quality of search results [1].

Web mining

Web mining can be defined as: Extract interested, useful patterns and implicit information from the WWW resources and behavior. In general, Web mining can be divided into three categories: Web content mining, Web structure mining and Web usage mining.

The classification of Web mining:

Web content mining is used to extract the text, image or other information and knowledge component of the web content. Search engines, intelligent agents, and some recommend use content mining to help the user in the vast network of space to find the necessary content. Web content mining has two strategies: page text mining; process results for search engine query further to get more accurate and useful information. Web structure mining is used to extract the network topology information, that is, the link between pages of information. Mine knowledge from the WWW organization and links. Which pages are linked to other pages? Which pages point to other page?

Web usage mining is used to extract about the customer how to use the browser and use the page links. It extracts interested patterns from the access to records of Web.

Forexample, which pages are the client accesses? How long spent on each page?

WWW. Each server retains the Web access log recording information for the user access and interaction. Analysis of these data can help understand the user's behavior, thus improving the structure of the site, or to provide users with personalized services.[2]

Semantic Web

The basic idea of Semantic Web[3] is that embed machine-readable, on behalf of certain types of knowledge mark in the Web message. So that the data in the Web is not only used to display, but also be understood by the machine so as to enhance the quality of the information services and explore a variety of new, intelligent information services. If the knowledge that reflect the link between data and application are embedded in a variety of different information sources in a user transparent manner, Web pages, database, procedures will be able to link up through the agent and each other collaborate. According to Berners-Lee's vision, the semantic network constituted by seven levels is constituted of a layered architecture [4]. As shown in following Table.

Layers	name	Description
Layer1	Unicode and URI	The Semantic Web-based Unicode Processing resources to encoding, URI (Uniform Resource Location negative Responsible for identification of resources
Layer2	XML+NS+XML Schema	Used to represent the data contents and structure
Layer3	RDF+RDF Schema	Used to describe resources on the Web and types
Layer4	Ontology Vocabulary	Describe the various types of resources and the relationship between resources
Layer5	Logic	On the following four layers operate on the basis of logical reasoning
Layer6	Proof	According to logic, to verify statements in order to draw conclusions
Layer7	Trust	The most element of a trust relationship between users

Technologies for Semantic web Mining

A. *XML –eXtensible Markup Language*

XML was designed as simple way to store or send documents across the web, which allows a developer to add meaning to the data being stored or transmitted. This functionality is made available by allowing a developer to create his or her own meaningful tags that contain data. When the XML file is then interpreted, a computer application can parse the tags and perform certain functions on that data as determined by the content attributes f the tags, which encloses it.

B. *Resource Description Framework (RDF)*

RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.

RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This graph view is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations.

The ResourceDescription Framework(RDF) is a family of World WideWeb Consortium (W3C) specifications [5] originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources.

C. *Web Ontologies*

Ontology refers to a document or file that formally defines the relation among terms. The typical web ontology consists of both taxonomy and a set of inference rules. The taxonomy defines all the classes of objects and any relationships between them. The use of classes, subclasses ad relations are very powerful tools to use over the web, because they allow developers to express large numbers of relations among different entities by assigning properties to classes and allowing subclasses to inherit these properties. The inference rules allow an application to make decisions based on the classes supplied without needing to actually understand any of the information provided. Furthermore, ontologies can be used to perform a variety of different functions other than simple deductions. Because more information is presented about a concept, they can act to improve the accuracy of search engine requests and allow applications to perform a wide variety of tasks autonomously.

D. *Agents*

Semantic Web lies with agents are the actual software applications that collect content from all over the web, process the information and exchange

the results with other software agents. These agents will provide the backbone to the semantic web, in that they will be able to exchange data with other agents even though the data is not specifically designed for the particular agent. Furthermore, these software agents are not only responsible for moving information backwards and forwards but also for exchanging digital signatures and proofs. The software agent can perform checks based on the RDF's triples and inference rules to ensure that the data it has received are accurate[6].

Supporting Tools and Techniques

A. Web Ontology Language (OWL)

OWL is intended to be used when the information contained in documents needs to be processed by applications, as opposed to situations where the content only needs to be presented to humans. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their interrelationships is called ontology. OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web. OWL is a revision of the DAML+OIL web ontology language incorporating lessons learned from the design and application of DAML+OIL[7].

B. SPARQL :

RDF is a directed, labeled graph data format for representing information in the Web. This specification defines the syntax and semantics of the SPARQL query language for RDF. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. SPARQL also supports extensible value testing and constraining queries by source RDF graph. The results of SPARQL queries can be results sets or RDF graphs[8].

C. Concept Description Language (CDL)

CDL(Concept Description Language) is a computer language to describe concept structure of content. CDL consists of a variety of family languages which are based on nested network data model and concept definition dictionary

D. Perl

Perl is a programming language developed by Larry Wall, especially designed for text processing. It stands for Practical Extraction and Report Language. It runs on a variety of platforms, such as Windows, Mac OS, and the various versions of UNIX. Perl is a general-purpose programming language originally developed for text manipulation and now used for a wide range of tasks including system administration, web development, network programming, GUI development, and more.

E. Support Vector Machines (SVM)

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. More formally, a support vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks.

F. Latent Semantic Analysis (LSA):

Latent Semantic Analysis(LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the totality of information about all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and set of words to each other. The adequacy of LSA's reflection of human knowledge has been established in a variety of ways[9].

G. Latent Semantic Indexing (LSI):

Latent Semantic Indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called Singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts.

Some Problems

Web content has to cater to two distinct needs: those of the human reader and those of the machine reader. Certain human oriented concepts, particularly abstract ones (e.g. love, hate, jealousy etc) are almost impossible

express in machine readable terms. Furthermore, concepts that apply in one situation are often not as applicable in other situations.

Word Sense Ambiguation is the another problem to define the proper sense or meaning to the ambiguous words.[10]

Conclusion &Future scope

The required information can effectively and accurately extracted and retrieve by using semantic index and ontology based approach. To handling semi-structured and unstructured data and defining rules in semantic web is also the great challenge.

Lot of work is done for domain specific retrieval,in future some domain independent and collaborative work is requires.

References

Z. Gyongyi and H. Garcia-Molina, "Web Spam Taxonomy," Proc. of the First International Workshop on Adversarial Information Retrieval on the Web, 2005.

Wen-Wei Chen, "Data Warehouse and Data Mining Tutorial,"[M], Beijing: Tsinghua University Press, 2006, 4

ZhongXue Ling, "Semantic Web in the core layer of technicalanalysis,"[M], South China Financial Computer ApplicationsTechnology, 2007,10

Ge Jian-Jiang, "Semantic Web principles and technology," [M],Beijing: Science Press, 2007, 3

http://www.w3.org/standards/techs/rdf, Created by RDF Working GroupPublication date: 2014-02-25

Berners-Lee T.,Hendler J. and LissilaO. 2001. The SemanticWeb[online]

http://sciamcom/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

http://www.w3.org/TR/owl-features , Editors: Deborah L. McGuinness, Frank van Harmelen, 2004 W3C

http://www.w3.org/TR/rdf-sparql-query, Editors: Eric Prud'hommeaux, Andy Seaborne, 2006-2007 W3C

Landauer, T. K., Foltz, P. W., &Laham, D. "Introduction to Latent Semantic Analysis", Discourse Processes, **25**, 259-284. (1998)

Wei Jan Lee and Edwin Mit, "Word Sense Disambiguation By Using Domain Knowledge", 2011 International Conference on Semantic Technology and Information Retrieval 28-29 June 2011, Putrajaya, Malaysia, IEEE 2011.

*****

CLOUD COMPUTING SECURITY ISSUES IN BIG DATA

Amruta Uphade, Pratik Ingle, S.E.Tayde

Department of Computer Science, S.S.S.K.R. Innani Mahavidyalaya, karanja lad

Amruta.uphade@outlook.com, Ingleprattik1@gmail.com, setayde_24@rediffmail.com

ABSTRACT

In this paper, we discuss security issues for cloud computing, Big data. The main focus is on security issues in cloud computing that are associated with big data. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries. Big data is a key concept that cannot be overlooked in the IT world considering the prominent increase in data, and data related services, it is important to explore this field and look at ways to improve data service delivery especially in the cloud. Cloud computing on the other hand helps in tackling the issue of storage and data service.

Keywords: *Cloud Computing, Big Data.*

Introduction

Many companies are using the technology to store and analyze pet bytes of data about their company, business and their customers. As a result, information classification becomes even more critical. For making big data secure, techniques such as encryption, logging, honey pot detection must be necessary. In many organizations, the deployment of big data for fraud detection is very attractive and useful.