A study on Semantic Web Mining

Nagina Yadav, Sanatan Sukeja
ITM University Gurgaon

Abstract

This paper is to define the combination of two scientific research areas Semantic Web and Web Mining is known as Semantic Web Mining. The amount of Semantic Web data is increased then it is good for researchers to apply data mining techniques on it. Semantic Web Mining is deal with very complex and heterogeneous data. In this paper we analyze and classify Web mining techniques which are applicable in different task of Semantic Web in form of an analytical framework.

1. Introduction

Semantic Web Mining is a combination of two important areas one is Semantic Web and other is Data Mining. Semantic Web is used to define a meaning to data, create complex and heterogeneous data structure. Data Mining is used to extract important and interesting patterns from same and less complex data. In a distributed manner the information environment, documents and objects has been joined together to interactive access. The most and important example of this environment is WWW. By using this WWW the users utilize hyperlinks and URL address for finding the useful information. At present time the web exceeds the ten billion pages, or more than six terabytes of data on three million servers. In a daily routine a million web pages are added on web, a typical page change in few month, and hundred gigabytes changes every month. Unstructured or Semi-Structured is the main problem of accessing, by this it is difficult to structure, standardize and organize, that level of complexity or problems in large volumes databases make them harder to managed and information retrieval is also impossible to managed. The method is to tackle this problem is Web Mining. Web Mining used data mining techniques to exploring and extracting the information automatically from documents and web services. We can divide the task of web mining into three types of categories: Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM). Semantic web is a good extension of the current web in which the information is well defined by meaning and changing in web content into a machine readable form which promote the quality and intelligence of the web. In Semantic Web explicit metadata is also used for accessing the information. Semantic Web is not a different entity from current Web. For creating and design a semantic web, the appropriate method is web mining. In this research paper we study about different applications of Web Mining for implementing semantic web. The role of introducing web mining approaches for constructing tasks of semantic web. This paper gives a general overview of semantic web.

2. Semantic Web

Semantic Web is used to provide meaning to the data from different types of web resources to allow the machine to interpret and understand the data to give answer and satisfy the web requestors. Semantic Web is a part of the second generation web (Web2.0) and its original idea derived from the World Wide Web consortium. Semantic Web represents the extension of the World Wide Web that gives users of Web and the ability to share their data and websites using the meaning of the web.

2.1. Semantic Web Representation Techniques

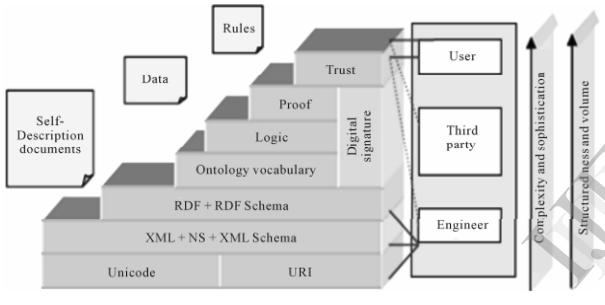
The techniques and models are used to define and represent the Semantic Web are W3c named Extensible Mark-up language (XML), RDF, Ontology language.

2.1.1. XML. XML is a good technique to store, organize and retrieve data on the web. It is enable users to create their own tags and allows them to define their content easily. By using this we represent a relationship among data and its Semantic Web.

2.1.2. RDF. RDF is an language that commonly used to enables the facility for storing resources and

Information that are available in the WWW (World Wide Web). RDF uses his own vocabularies for domains. Elements contained in RDF are of three types: Resource, Literals, and Properties. Resources are those in which entities are identified by URLs. Literals used atomic values like strings and numbers. Properties used binary relationship identify by URIs. RDF is a way to represent the data that could be defined on the web in a different manner.

2.1.3. Web Ontology Language. OWL is a more complex language with better machine interpretability then RDF. It is used to identify the resources nature and their relationship. The architecture of semantic web is based on 7-layers on the vision of Sir Berner's lee: 1.URI, 2.XML, NS, and XML schema, 3.RDF and RDF Schema, 4. the ontology vocabulary, 5.logic, 6.proof, 7.trust.



Layers:

i. **URI**: it is in charge of encoding process and its identification.
ii. **XML, NS, and XML schema**: This layer is in charge of
1) The separation of data content, data structure, and the performance format based on linguistic
2) Representing them to use a format language.
iii. **RDF and RDF schema**: Used to define the information on WWW and its type using a semantic model.
iv. **Ontology vocabulary**: This layer is concentrated on semantics of information by defining the knowledge shared and the semantic relations within different sorts of information.
v. **Logic**: It takes the responsibility of providing the intelligent services such as logical reasoning by supply axioms and inference principles.
vi. 6. **Proof and Trust**: These two layers deal with enhancing the security of web by using encryption and digital signature. This architecture of Semantic Web is shown in diagram.

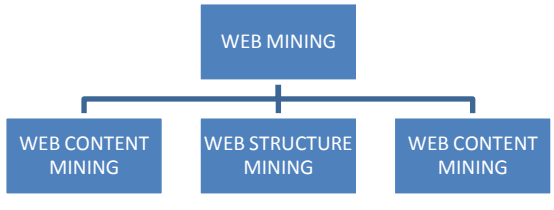
3. Data Mining: Web Mining

Data mining plays a great role in the development of many domains. One of the widely used techniques of data mining is association rule mining which define the algorithm of generating patterns in the form of X=>Y in each X and Y are non empty subsets of items and this rule has a support and confidence value more than the threshold of users. Web Mining is the combination of both Text and Data Mining to mine the biggest information resource (web). The process of Web Mining is divided into these subtasks:

i. Resource finding: Retrieving Web documents.
ii. Information selection and preprocessing: From retrieved Web resources we automatically select and pre-processing information.
iii. Generalization: Discovers patterns at individual Web sites as well as across multiple sites.
iv. Analysis: Validation and verification of patterns that are mined.

3.1. Types of Web Mining

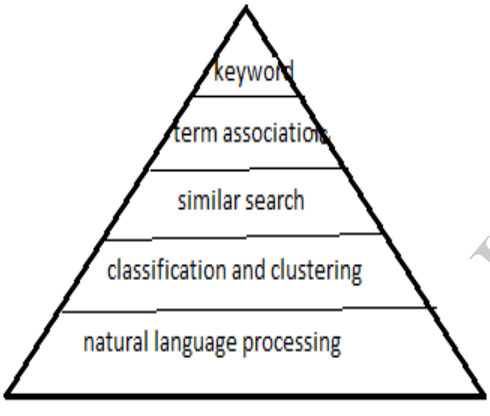
i. Web Content Mining
ii. Web Structure Mining
iii. Web Usage Mining



3.1.1. Web Content Mining. Web content mining is concentrating on information and knowledge to mine from different contents such as text and image. For help and improve result of areas like search engines to deliver precise and beneficial information to the users of web. Web content is consists of many types of data such as image, text, audio, video, data about data and hyperlinks. Web content data is consist of unstructured Data such as free text, Semi-structured data such as HTML documents, and structured data such as Tables and database generate HTML pages. The goal of Web-content mining is to assist or to improve the Information finding or filtering the information. To build a new model of data on the Web, a more sophisticated queries other than the keywords based Search could be asked. Mostly search engine are based on keyword Searching. WCM is used the basic IR technology. Web Mining divides the web content

mining into agent-based and database approaches. The agent based approaches have software systems (agents) that perform the content mining. For example, they may use user profile and or knowledge domains. Personalized web agents use information about user preference to direct their search. The Database approaches used the web data belonging to database. Content Mining is also called text mining. Text mining functions can be viewed as a hierarchy with the simplest function at the top and the complex at the bottom. Research is currently going on the use of natural language processing techniques in content or text mining to uncover the hidden semantics, like as question and answers system. The traditional mining operations are keyword search, similar measures, clustering and classification.

3.1. Types of Web Mining

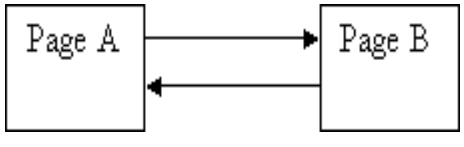


3.1.2. Web Structure Mining. It tries to recognise the topology information of network, mining connection between the Web pages. It is used to discover the link structure on the Web. It is based on the topology of the hyperlinks with or without describing a links. It can be used to categorize Web pages and it is useful to generate information such as the similarity relationship among Web sites, documents. To improve the effectiveness of search engines and crawlers we used two algorithms one is page rank algorithm and another is hits algorithm.

PageRank: Page Rank is a algorithm used by Google. The Page Rank technique is designed to increase the effectiveness of search engines and improve efficiency. It is also used to measure the importance of a page return from a traditional search engine using keyword searching. The Page Rank value for a page is calculated based on the no of pages that point to that page. A measure is based on the number of back links to a page. A back link is a link pointing to a page rather than pointing out from a page. PageRank is an excellent method to prioritize the result of keyword searches. Page Rank is also help in full text searches in the Google system. We assume page A has pages N1……Nn point to it. The parameter d is a damping factor can be set between 0 and 1. Commonly set d as 0.85. C(A) is used as the no of links going out of page A. The Page Rank of page A:

PR (A) =(1-d) +d(PR(N1)/C(N1)+…..+PR(Nn)/C(Nn))

Page Rank or PR (A) used iterative algorithm for calculation. Page rank of a page is calculated as the sum of all the incoming links divided by the its outgoing links. The two main characteristics of search results ranking of PageRank algorithm: (1) Sorting is independent of search keywords (2) Ranking is independent of the specific content of the page. Page Rank is a static algorithm and has nothing to do with the query. PR values of pages are calculated by off-line, which reduce the computation time and response time when searching is done online. We take the simple example: two pages, each linking to the other:



Every page has one outgoing link (the outgoing value is 1, i.e. D(A) = 1 and D(B) = 1).

We take a guess at 1.0 and do calculations:

d = 0.85

PR(A) = (1 – d) + d(PR(B)/1)

PR(B) = (1 – d) + d(PR(A)/1)

Put value into the above formulas:

PR(A) = 0.15 + 0.85 * 1
= 1

PR(B) = 0.15 + 0.85 * 1
= 1

Implementation of Page Rank

Data to implement:
*Only single line comments are used.
* Every line has its own rule.
* We only used (:) or space.
* To terminate the statement we don't used ;
* To separate out bounds we don't used ,
* Every page is used a separated line.
* White spaces are ignored.
* Page: out bounds (separated by spaces)
1 : 3 5 4
2 : 1 8 3
3 : 7 2
4 : 9 2
5 : 8 4
6 : 1 5 2 7 3
8 : 5 3 6 9 8
9 : 1 6

Output

Page Rank

1 : 0.11128192693223575

2 : 0.11148041620908788

3 : 0.1113236703206586

4 : 0.11104452602105681

5 : 0.11092670382696641

6 : 0.11088581064152737

7 : 0.11088729911839655

8 : 0.11128299145676111

9 : 0.11088665547330939

Sum: 1.0 (must be approximately equal to 1)

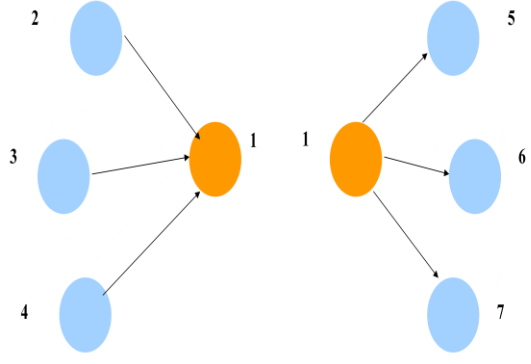
•

6.2. Types of pages

HITS algorithm used two types of pages first is Authority and second is Hub.
Authority: Page that provide an important and trustworthy Information on a given topic. A better authority is pointed by many goods hubs.

Hub: Pages that contain link to authorities. A better hub points by many good authorities.

Diagram
Authority and hub



$$a(1) = h(2) + h(3) + h(4)$$

$$h(1) = a(5) + a(6) + a(7)$$

In degree: Number of incoming links to a given node. It is used to measure authority.

Out degree: Number of outgoing links from a given node here it is used to measure the hubness. Hub and authority together make a bipartite graph. HITS are a General algorithm used to calculate the authority and hubs in order to rank the data**.**

Algorithm:
There are two basic steps on which the HITS algorithm performs a list of iterations**.**

Authority Update: It is used to update each node's *Authority value* to be equal to the total sum of the *Hub value* of each node that points to it. A node is given a high authority number by being linked to by pages that are initialized as Hubs for information.
Authority Update Rule

$\forall p$, we update $\text{auth}(p)$ to be the summation:

$$\sum_{i=1}^{n} \text{hub}(i)$$

Where the total number of n pages connected to p pages and a page I am connected to

p. So the Authority value of a page is the total sum of all the Hub value of pages that point to it.

Hub Update: It is used to update each node's *Hub value* to be equal to the total sum of the *Authority value* of each node that it points to. Any node is given a high hub value by linking to nodes that are selected to be authorities on the subject.

Hub Update Rule

$\forall p$, we update $\mathrm{hub}(p)$ to be the summation:

$$\sum_{i=1}^{n} \mathrm{auth}(i)$$

The Hub value and Authority value for a node is calculated with the following algorithm steps:

i. Start with every node has a hub value and authority value of 1.
ii. Let the Authority Update Rule to be run
iii. Let the Hub Update Rule to be run.
iv. Normalizing the values by dividing every Hub value by square root of the total sum of the squares of all Hub value, and dividing each Authority value by square root of the total sum of the squares of all Authority value.
v. Repeat the rule given in the second step as required.

3.1.3. Web Usage Mining. It is focus on the discovery of the information that users of web are searching .this is done by mining web users record to gain information about the use of browser and links of page that helping in understanding user's behaviour and personalize web services.

Taxonomy of web usage mining:

1. Personalization for a user can be achieved by keeping track of previous pages that are accessed.
2. To determine frequent access we need to identify links to improve the performance for future access.
3. Frequently accessed information can be used for caching.
4. Gathering of business intelligence to improve sales and advertisement we used web usage patterns.

Web usage mining consists of following types of activities:

1. Pre-processing: It is used to reformat the web data before processing.
2. Pattern discovery: It is used to form the main part of mining activities. Activities are used to find the hidden patterns for log data.
3. Pattern analysis: Result of the discovery activities is looked and interpreting by pattern analysis.

4. Conclusion

Semantic Web Mining is a new and fast-developing research area by combining Web Mining and Semantic. The merging of both areas using semantic structures in the Web to reach the results of Web Mining and to make the Semantic Web by deploying the Web Mining techniques. Semantic web is the main domain of research area. In this paper we analyze and classify Web Mining techniques which are applicable in different task of Semantic Web in form of an analytical framework.

5. References

[1] Graph Theory Book with application to engineering and computer science by Narshing Deo.
[2] Data mining book by Han and Kamber
.[3] Hamed Hassanzadeh and Mohammad Reza Keyvanpour "Semantic Web Requirements through Web Mining Techniques" International Journal of Computer Theory in 2012.
[4] Qudamah K. Quboa, Mohammad Saraee "A State-of-the-Art Survey on Semantic Web Mining".
[5] Sergey Brin and Lawrence Page "The Anatomy of a large scale hyper textual web search engine by Stanford University.
[6] Marc Najork Compairing the Effectiveness of HITS and SALSA.
[7] Taher havliwala, Sepandar kamvar Glen Jeh An analytical comparison of Approaches to Personalizing page rank.