Globus An International Journal of Management & IT
A Refereed Research Journal
Vol 8 / No 1 / Jul-Dec 2016 ISSN: 0975-721X

A STUDY ON WEB MINING AS WEB SERVICES FRAMEWORK

*Dr Pankaj Saxena

Abstract

The World Wide Web has turned to be one of the largest information sources. It is heterogeneous, explosive, dynamic and mostly unstructured data repository. Some companies use the Web to find out more about their competition. Another example in the academic context is teams involved in research that use the Web to find out about related work. Students may use the Web to find out more about their University, certain departments or particular lectures and projects. Some individuals use it to keep up to date with the latest news, according to their interests. Because Web supplies extensive amount of information it raises the complexity of how to deal with the information from the different point of view. The business experts want to have tools to learn the customer's needs. The users want to have the efficient search tools to find significant information easily. All of them are expecting tools or techniques to help them satisfy their demands and/or solve the problems encountered on the Web. Therefore, Web Intelligence is required to help organizations in decision making as well as also help users in finding relevant information.

Keywords: World Wide Web, Web Mining.

Introduction

Web content mining focuses mainly on the structure within a document i.e. inner document level. Web documents can be in different formats. They can be HTML (Hyper Text Markup Language) documents or XML (Extensible Markup Language) documents, etc.

HTML documents contain semi-structured or more structured data like the data in the tables which are generated by the databases or automatically generated, but most of the data is unstructured free text data. For example, one web site could display the price of same car in numeric figure others may do in words. An attribute may have an atomic value in one document, but a set of values in other documents. Different attributes in documents may have semantically similar meaning across WWW or vice versa. HTML was firstly implemented by Tim Berners-Lee at CERN, and became popular by the Mosaic web browser developed at NCSA. HTML instructions divide the text of a web page into sub blocks called elements. The HTML elements can be examined in two ways. One way is that define how the body of the web document is to be displayed by the web browser, and another way is that define the information about the web document, such as its title or its relationships to other web documents.

XML documents contain structured information which contains both the content and the information about what content includes and stands for. Almost all documents have some structure. XML has been accepted as a markup language, which is a mechanism to identify structures in a document. XML specification determines a standard way to add markup to documents. XML doesn't specify semantic or tag set. In fact it is a meta-language for describing markups. It provides mechanism to define tags and the structural relationships. All of the semantics of an XML document will either be defined by the applications that process them or by style sheets. The design goals of XML emphasize simplicity, generality, and usability over the Internet.

Dynamic server pages are also important part of web content data. Dynamic content can be any web content, which is processed or compiled by the web server before sending the results to the web browser. On the other hand, static content is content, which is sent to the browser without modification. Common forms of dynamic content

Associate Professor, R.B.S. Management Technical Campus, Agra, India.

are Active Server Pages (ASP), Pre-Hypertext Processor (PHP) pages and Java Server Pages (JSP). Today, several web servers support more than one type of active server pages.

Web content mining uses data mining techniques; it differs from data mining because Web content data are mostly unstructured and/or semi-structured, while data mining deals mainly with structured data. a large number of multimedia sources are contained on the web, such as images, videos, audios, and animations, which make multimedia content indexing much more difficult than document indexing, prompting the need for "multimedia web mining". Multimedia web mining is related to mining operations on multimedia files on the web for the extraction of knowledge embedded in multimedia web data. The semantic web is based on "ontologies", which are metadata related to web page content that make the site meaningful to search engines. Research activities in this field also involve using techniques from AI such as Information Retrieval [IR], Natural Language Processing [NLP], and Image processing and computer vision.

Review of Literature

The paper discusses various approaches for web service composition. Further the paper provides a strong basis for creating reference models for web service composition. In the first part state of art is discussed and in the later part of the paper (Messaoud W. B., Ghedira K., Halima Y. B., & Ghezala H. B., 2015), framework design of the reference model is discussed.

The man purpose of the paper is to study systematic review of the literature done so far till 2015 and to develop a classification on QoS aware web service composition. Paper is a thorough analysis of different databases that focuses on meta- heuristic algorithms for QoS aware web service composition. Paper says QoS aware web service composition is NP- Hard and it needs to be solved in acceptable amount of time. After getting data from different research articles it's been observed that most widely used meta-heuristic techniques are Genetic algorithm (38%), PSO (22%) and ACO (10%). The paper saves time of various researchers in the field of web service composing by giving a thorough overview of web service composition.

The paper is graph based and it shows an approach to integrate the services or we can say it shows a different manner of composition of web services through discovery and matchmaking (Rodriguez-Mier P., Pedrinaci C., Lama M., & Mucientes M., 2016).

The study uses semantic context-aware planning graph and proposes a semantic adaptable web services composition method. There are 3 stages used in composition.

- Construction of planning graph based on discovery and selection of atomic web services which takes users context and his preferences. Semantic relation between input and outputs of web service are used to enhance planning graph.
- Backward search is used in which they are extracting a set of best composed web services.
- Based on semantic and context scores, ranking is provided to set of extracted solutions.

Web Mining

Web mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires creative applications of data mining and/or text mining techniques and also its own unique approaches. In the past few years, there was a rapid expansion of activities in the Web content mining area. This is not surprising because of the phenomenal growth of the Web contents and significant economic benefit of such mining. However, due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems. In this tutorial, we will examine the following important Web content mining problems and discuss existing techniques for solving these problems. Web mining under web services is a rapid growing research area. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. Web content mining aims to extract/mine useful information or knowledge from web page contents.

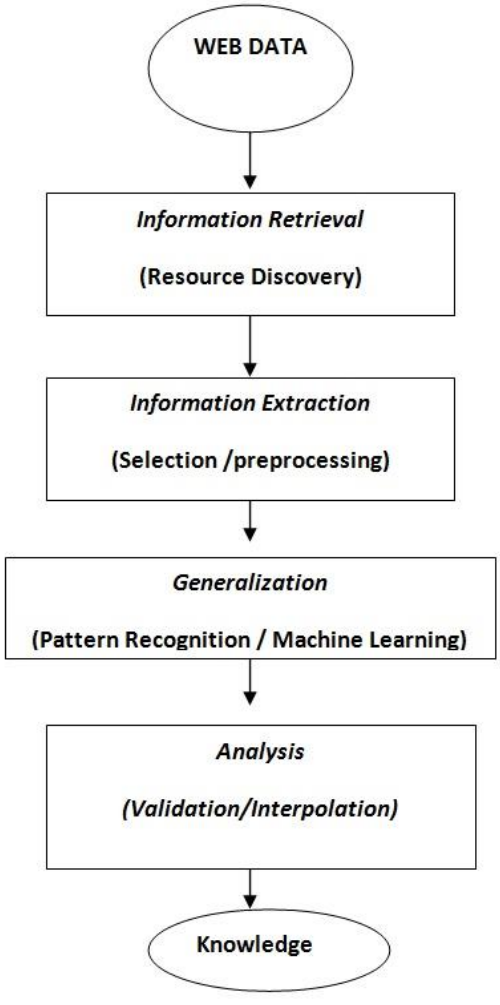

Figure 1: Web Mining Process

Opportunities and Challenges of Web Services

The Web-an immense and dynamic collection of pages that includes countless hyperlinks and huge volumes of access and usage information-provides a rich and unprecedented data mining source. However, the Web also poses several opportunity and challenges to effective resource and knowledge discovery:

- The coverage of Web information is very wide and diverse. One can find information about almost anything. The Web serves a broad spectrum of user communities. The Internet's rapidly expanding user community connects millions of workstations. These users have markedly different backgrounds, interests, and usage purposes. Many lack of good knowledge of the information network's structure, are unaware of a particular search's heavy cost, frequently get lost within the Web's ocean of information, and one can lengthy wait required to retrieve search results.
- Information/data of almost all types exist on the Web, e.g., un-structured, semi-structured structured data like texts, multimedia data, tables etc. Much of the Web information is semi-structured due to the nested structure of HTML code.
- Much of the Web information is redundant. The same piece of information or its variants may appear in many pages.
- Only a small portion of the Web's pages contain truly relevant or useful information. A given user generally focuses on only a tiny portion of the Web, dismissing the rest as uninteresting data that serves only to swamp the desired search results.
- Much of the Web information is linked. There are hyperlinks among pages within a site, and across different sites.
- The Web is noisy. A Web page typically contains a mixture of many kinds of information, e.g., main contents, advertisements, navigation panels, copyright notices, etc.
- The Web is also about services. Many Web sites and pages enable people to perform operations with input parameters, i.e., they provide services.
- The Web constitutes a highly dynamic information source. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues. Not only does the Web continue to grow rapidly, the information it holds also receives constant updates. News, stock market, service center, and corporate sites revise their Web pages regularly. Linkage information and access records also undergo frequent updates.
- Web page complexity far exceeds the complexity of any traditional text document collection. Although the Web functions as a huge digital library, the pages themselves lack a Uniform structure and contain far more authoring style and content variations than any set of books or traditional text-based documents. Moreover, the tremendous number of documents in this digital library has not been indexed, which makes searching the data it contains extremely difficult.
- Although keyword, address, and topic-based Web search engines already support information searches, web mining will play an important role in Web intelligence because the Web's current incarnation still cannot provide high-quality, intelligent services. Several

factors contribute to this problem and motivate our research.

- Databases provide high-quality, well-maintained information, but are not effectively accessible. Because current Web crawlers cannot query these databases, the data they contain remains invisible to traditional search engines. Conceptually, the deep Web provides an extremely large collection of autonomous and heterogeneous databases, each supporting specific query interfaces with different schema and query constraints. To effectively access the deep Web, we must integrate these databases.
- Because current Web searches rely on keyword- based indices, not the actual data the Web pages contain, search engines provide only limited support for multidimensional Web information analysis and data mining.

All these tasks present major research challenges and their solutions also have immediate real-life applications. These challenges have promoted research into efficiently and effectively discovering and using Internet resources, a quest in which web mining will play an important role regarding Web Intelligence.

The Web constitutes a highly dynamic information source which is in large size and has a great Complexity. WI may be viewed as an enhancement or an extension of Artificial Intelligence and Information Technology. WI may become a sub-area of Artificial Intelligence and Information Technology or a child of a successful marriage of AI and IT. Web Intelligence (WI) concerns the design and implementation of intelligent web information systems on the new platform of the Web and Internet Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. The following few web-based tools of Web Intelligence to form intelligent web information systems are:

- Web Information System Environment and Foundations:
- Web Human-Media Engineering
- Web Information Management
- Web Information Retrieval
- Web-Based Applications
- Web Agents
- Web Mining and Farming

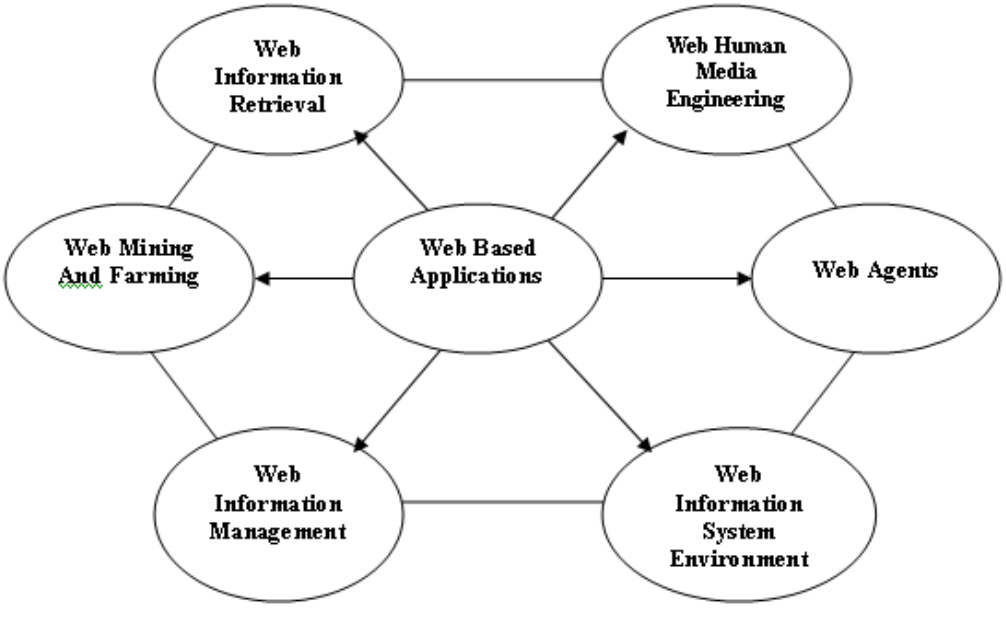

Figure 2: Different Web-Based Tools of Web Intelligence

Conclusion

The some of the more popular industries which are getting the benefits of web mining are Ecommerce & Retail, Travel & Tourism, Automobiles, Telecommunications, Media & Entertainment, Information Technology, Computers & Equipment, Jobs & Recruiting, and Financial Academic Research. But still, the Web present new challenging and research problems as discussed above. There is a requirement of existing theories

and technologies of web mining need to be modified or enhanced.

To use the large amounts of information efficiently on the Web to make the information processing intelligent, personalized and automatic is the most important applications of the current web mining technology. The main areas, Web mining covers for Intelligent Web Information System, in case of Web Intelligence are data mining and knowledge discovery, hypertext analysis and transformation, learning user profiles, multimedia data mining, regularities in Web surfing and Internet congestion, text mining, web-based ontology engineering, web-based reverse engineering, web farming, web-log mining, web warehousing.

This research work research work primarily focuses on web mining sub tasks, web mining classification, Web mining software tools and various web mining techniques to improve Web Intelligence. Researchers may clean, condense, and transform this research work to retrieve and analyze significant and useful information to determine best possible improvements of web mining techniques and effective software tools which can help to improve web intelligence to form the Intelligent Web Information Systems.

References

1. Messaoud W. B., Ghedira K., Halima Y. B., & Ghezala H. B. (2015). "Survey of web service composition", 5th International Conference on Information & Communication Technology and Accessibility (ICTA), Marrakech, pp. 1-7.
2. Rodriguez-Mier P., Pedrinaci C., Lama M., & Mucientes M. (2016). "An Integrated Semantic Web Service Discovery and Composition Framework", in IEEE Transactions on Services Computing, vol. 9, no. 4, pp. 537-550.
3. Rawat, Deepak (2015). "CMS Deployment and Development", Globus An International Journal of Management & IT, Vol 7, No 1, pp 32-33.
4. W., Kim (2013). "Cloud Computing: Today and Tomorrow", Journal of Object Technology, Vol. 8, No.1, pp. 65-72.
5. Kumar, Dharanikota Kiran and Dawra, Dr. Sudhir (2015). "Existing Evaluation Methods of Software Process Models", Globus An International Journal of Management & IT, Vol 7, No 1, pp 45-48.
6. Athena (2009). "Cloud Computing Distributed Internet Computing for IT and Scientific Research", Journal of IEEE Internet Computing, Vol. 13, Issue 5, pp. 10-13.