

Mining Paradigms: The Roadmap For Semantic Web Mining

MPS Bhatia,

Akshi Kumar

Netaji Subhash Institute of Technology, Delhi University

mpsbbhatia@nsit.ac.in, akshi.kumar@gmail.com

Abstract --World Wide Web (also known as Web) is a glaring & interactive medium to promulgate information today. The recent surveys claim that 85% of internet users use search engines and search services to find specific information. The same surveys, however, show that users are not satisfied with the performance of the current generation search engines. The slow retrieval speed, communication delays, and poor quality of retrieved results are commonly cited glitches. We discuss the development of new techniques targeted to resolve some of the problems associated with Web-based Information Retrieval. This paper expounds the roadmap for Semantic Web Mining by investigating the evolution of mining models from the basic Information Retrieval & Extraction framework to the Data Mining, Web Mining and Semantic-Web Mining paradigms.

Keywords --Information Retrieval & Extraction, Data Mining, Web Mining, Semantic Web, Semantic Web Mining.

1. INTRODUCTION

The World Wide Web is today's largest warehouse of knowledge. It is a huge, widely distributed, global source for information services, hyper-link information, access and usage information and web-site contents & organizations. With the transformation of the Web into a ubiquitous tool for .e-activities. such as e-commerce, e-learning, e-government, e-science, its use has pervaded to the realms of day-to-day work, information retrieval and

business management. In addition, this renovation has refurbished as well as revolutionized the role and availability of information. By all measures, the Web is enormous and growing at a staggering rate, which has made it increasingly intricate and crucial for both people and programs to have quick and accurate access to Web information and services. Buried in the enormous, multi-dimensional, heterogeneous and distributed information on the Web is the knowledge having great potential value. With the rapid development of the Web, it is imperative to provide users with tools for efficient and effective resource and knowledge discovery. Search engines have assumed a central role in the World Wide Web's infrastructure as its scale and impact have escalated. Although the web search engine assists resource discovery, it is far from satisfying for its poor precision and recall. [1, 2, 3]. Stakeholders could encounter, among others, the following problems when interacting with the Web:

a) *The "Abundance" problem:* - With the phenomenal growth of the Web, there is an ever increasing volume of data and information published in numerous Web pages.

b) *Web Search results usually have low precision & recall:*- For finding relevant information, the search service is generally a keyword-based, query-triggered process which results in problems of Low Precision (difficulty to find relevant information) & Low Recall (inability to index all information available on the web).

c) *Limited query interface based on keyword-oriented search:* - It is hard to extract useful knowledge out of information available because the search service used to find out specific information on the Web is retrieval-

oriented, whereas to extract potentially useful knowledge out of it, is a data-mining oriented, data-triggered process.

d) *Lack of Personalization of Information & Limited Customization to individual users*:-Most knowledge on Web is presented as natural-language text with occasional pictures and graphics. This is convenient for human users to read and view but difficult for computers to understand. It also limits the state of art search engines, since they cannot infer meaning. For example the occurrence of word ‘bat’ refers to a bird or to a cricket bat. These factors uphold the inevitable creation of intelligent server and client-side systems that can effectively mine for knowledge both across the Internet and in particular web localities.

This paper expounds the roadmap for Semantic Web Mining. It tracks and investigates the evolution of mining models from the basic Information Retrieval & Extraction to the Semantic-Web Mining paradigm. The taxonomy related to the entire discovery framework including the issues and process of Information Retrieval & Extraction, Data Mining, Web Mining, Semantic Web, and Semantic Web Mining have been summarized and the relationships between them have been illustrated.

II. THE TRIPLET OF DATA, INFORMATION & KNOWLEDGE

Data is an expression of feedback; a statement (rightly or wrongly so) about an observation.

Information is contextualized data and knowledge is a phenomenon that implies our ability to use the information for reasoning and decision making, i.e., it is the basis of what you can, will, would, should or might do with information.

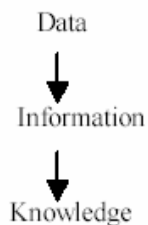


Figure1: The triplet

III. DISCOVERY FRAMEWORK TAXONOMY

A. *Data Mining*:-Data mining has been defined as "the non-trivial extraction of implicit, previously unknown, and potentially useful information from large data sets or databases. [4]. It is used to identify valid, novel, potentially and ultimately understandable pattern from data collection in database community.

B. *Knowledge Discovery*: - Knowledge discovery is the process of finding novel, interesting, and useful patterns in data. Data mining is a subset of knowledge discovery. It lets the data suggest new hypotheses to test. Thus, data mining is also known as Knowledge Discovery in Databases (KDD) [7].

C. *Information Retrieval (IR)*:-Automatic retrieval of all relevant documents while at the same time retrieving as few of the non-relevant as possible. It has the primary goals of indexing text and searching for useful documents in a collection. [1, 2, 5, 6, 7]. There are various ways to measure how well the retrieved information matches the intended information:

Precision

The proportion of retrieved and relevant documents to all the documents retrieved:

$$\text{precision} = \frac{| \{ \text{relevant documents} \} \cap \{ \text{retrieved documents} \} |}{| \{ \text{retrieved documents} \} |}$$

Recall

The proportion of relevant documents that are retrieved, out of all relevant documents available:

$$\text{recall} = \frac{| \{ \text{relevant documents} \} \cap \{ \text{retrieved documents} \} |}{| \{ \text{relevant documents} \} |}$$

Fall-Out

The probability to find an irrelevant among the retrieved documents:

$$\text{fall-out} = \frac{| \{ \text{irrelevant documents} \} \cap \{ \text{retrieved documents} \} |}{| \{ \text{retrieved documents} \} |}$$

D. Information Extraction (IE):-Information Extraction has the goal of transforming a collection of documents, usually with the help of an IR system, into information that is more readily digested and analyzed. [It includes the following subtasks: [1, 2, 5, 6, 7]

a.. Named entity recognition (NER)

Recognizing relevant entities in text.

b. Relation extraction

Linking recognized entities having particular relevant relations.

Applications of IE include:

- It can make Information Retrieval more precise.
- Summarization of documents in well defined subject areas.
- Automatic generation of databases from text.

E.. Difference between IR and IE

<i>Information Retrieval (IR)</i>	<i>Information Extraction (IE)</i>
Aims to select relevant documents, i.e, it finds documents.	Aims to extract relevant facts from the documents, i.e, it extracts information.
Views text as a bag of unordered words.	Interested in structure and representation of the document.

Table 1: IR vs. IE

Thus IE works at a finer granularity level than IR does on documents.

F. Web Mining: - Web mining refers to the use of data mining techniques to automatically retrieve, extract and evaluate (generalize/analyze) information for knowledge discovery from web documents and services. It is about making implicit or "hidden" knowledge explicit. The digital revolution and the phenomenal growth of the Web have lead to the generation and storage of huge amounts of data, prompting the need for intelligent analysis methodologies to discover useful knowledge from it. Due to the heterogeneous, semi-structured, distributed, time-varying and multi-dimensional facets of web data, automated discovery of targeted knowledge is a challenging task. It calls for novel methods that draw from a wide range of patent areas of Data Mining, Machine Learning, Information Retrieval, Natural Language Processing, Multimedia, and Statistics.

a. From Data Mining to Web Mining

Two notable and active areas of current research are Data mining and the World Wide Web. An expected alliance of the two areas, sometimes referred to as Web mining, has been the focus of several recent research projects and papers. Nevertheless, Web mining has many unique characteristics compared with data mining. Firstly, the source of Web mining is web documents. We consider the use of the Web as a middleware in mining database and the mining of logs & user profiles on the Web server still belong to the category of traditional data mining. Secondly, the Web is a directed-graph consists of document nodes and hyperlinks. Therefore, the pattern identified can be possibly about the content of documents or about the structure of the Web. Moreover, the Web documents are semi-structured or non-structured with little machine-readable semantic while the source of data mining is

confined to the structural data in database. As a result, some traditional data mining methods are not applicable to Web mining. Even if applicable, they must be based on the preprocessing of documents.

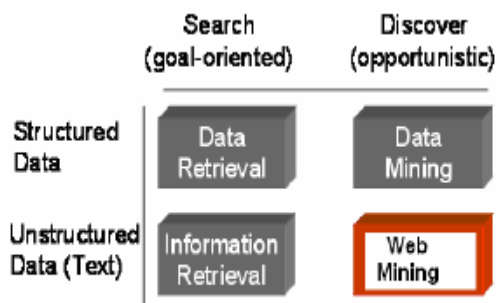


Figure 2: Web Mining View

b.. The Web Mining Process

Information Retrieval, Information Extraction and Web mining all have well-defined different goals. Web mining is a step ahead of Information Retrieval & Extraction. It does not intend to replace either of the two; instead the three technologies supplement each other. Each has its strong points and applications in point. On the other hand, Web mining can be utilized to increase the precision of Information Retrieval and improve organization of retrieval results as well, and Information Extraction can be used to improve the indexing part of the IR process. Thus, we can view IE as a preprocessing stage in the Web mining process, which is a step after the IR process and before the data mining techniques performed.

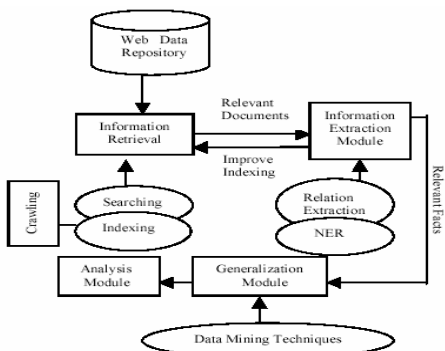


Figure3: The Web Mining Framework

The figure 3 above shows Web Mining as a derivative of the Data Mining process. The major components of Web

Mining Process include (i) Information Retrieval, (ii) Information Extraction, (iii) Generalization, and (iv) Analysis. [5, 7, 8, 9, 11]

3.6.2.1. Information Retrieval / Resource Discovery / Knowledge Crawling Module

The IR module automates the retrieval of relevant documents, using document indexing and search engines. It is done by web search and meta-search engines, or by crawlers. These approaches focus on a one-time analysis of websites and cannot deal with constantly changing web sites, such as news sites where the information is constantly added or modified.

- Locating document and services on the Web.
- Search engines: Yahoo!, Google, AltaVista.
- Retrieving data that is online or offline from the text sources available on the Web, E.g. electronic newsletter, electronic newswire, newsgroups, etc.
- Identifying sources that originally were not accessible from the Web but are accessible now.

3.6.2.2. Information Extraction Module

The IE module helps to identify document fragments that constitute the semantic core of the web. *Preprocessing* consists of two tasks: selecting interesting data from the downloaded web documents, and transforming this data into a formal representation. Most methods use wrappers for extracting simple data (e.g. proper names, prices, phone numbers, e-mail addresses, etc.) from web documents, and construct tables as formal representations

- Extracting specific information from Web resources.

- Removing stop words, stemming, finding key phrases, transforming representation
- Using wrappers (extraction patterns) to access the resource and parse its response.

3.6.2.3. Generalization Module

Generalization is the automatic discovery of patterns across multiple web documents. Most methods use data mining techniques for discovering association rules, clusters and classification trees and rules. It uncovers patterns at individual Web sites and across multiple sites and relates to aspects from pattern recognition/machine learning, and utilizes clustering and association rule mining.

3.6.2.4. Analysis Module

Analysis is validation and/or interpretation of the mined patterns. It corresponds to the extraction, interpretation, validation and visualization of the knowledge obtained from the web.

3.6.3. Web Mining Taxonomy

The diversity of information on the Web leads to the variety of Web mining, defined in the following three categories: [2, 9]

3.6.3.1. Web Usage Mining

- **DEFINITION:** Web Usage Mining refers to discovering user access patterns from Web usage logs.
- **EXAMPLE:** Web server access logs, Proxy server logs, Browser logs, User profiles, Registration data, User sessions or transactions, Cookies, User queries, Bookmark data, Mouse clicks and scrolls, any other interaction data.

- **APPLICATIONS:** Learning user profiles, Identifying associate terms, Web traffic control.

3.6.3.2. Web Structure Mining

- **DEFINITION:** Web Structure Mining refers to inferring useful knowledge from the structure of hyperlinks (in-links and out-links).

- **APPLICATIONS:** Webpage ranking in Google.

Web structure mining can be further divided into external structure (hyperlink between web page) mining, internal structure (of a web page) mining and URL mining.

3.6.3.3. Web Content Mining

- **DEFINITION:** Web Content Mining refers to extracting use information and knowledge from content in Web pages.

- **APPLICATION:** Document categorization, Sentiment classification.

Web content mining can be divided into text mining (including text file, HTML document, etc.) multimedia mining. Web mining techniques may use a combination of the above categories of techniques.

3.6.4. Application areas/systems

Information Retrieval, including Multimedia IR, Search Engines and Information Agents, Digital Libraries for Web Information, Content and Knowledge Management, Filtering and Recommender Systems, Commerce and Shopping Agents, Web Warehousing, Business Intelligence.

IV. SEMANTIC WEB

The Semantic Web is about developing standards to make knowledge machine-understandable, to allow Web agents to become more "intelligent", and ultimately to better support human-computer cooperation. It has opened a new window

to a plethora of applications and systems that takes benefit of machine-understandable information. [13, 14] Knowledge and the Web encompass a number of questions that arise in this situation:

- *Web --> knowledge*: How can knowledge be made accessible? In particular, how can implicit knowledge be turned into explicit knowledge?
- *Knowledge --> Web*: What structures can be given to or added to human-understandable knowledge to make it machine-understandable?

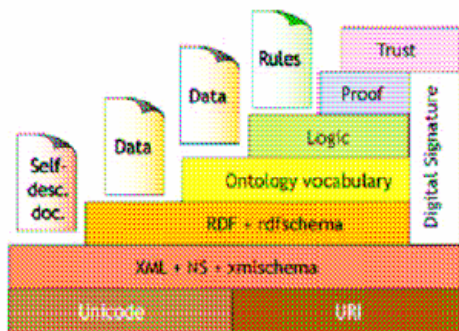


Figure4: The Semantic Web Framework[13]
[www.w3.org]

A. Where is semantic web today?

During the last few years many Semantic Web related technologies have emerged or have been elaborated.[16]

a) The status of ontology development languages is more stable now. The World Wide Web Consortium (W3C) who has been working intensively on semantic standards has approved the Resource Definition Framework (RDF) and the OWL Web Ontology Language (OWL) and hence provided a solid base to establish enterprise semantic applications and has implied a significant leverage of the Semantic Web from a research level to an industry standard for building next generation applications.

b) At the same time there have been evident motivations toward creating semantic-contents and more importantly the development of domain-specific ontologies. As a consequence Semantic Web has been widely accepted in the information technology branch with many research projects and industrial applications arisen from it.

V. SEMANTIC WEB MINING

Semantic Web Mining aims at combining the two emergent research areas of Semantic Web and Web Mining. [12, 15] The idea is to improve the results of Web Mining by exploiting the new semantic structures in the Web; and on the other hand, make use of Web Mining, for building up the Semantic Web by extracting useful patterns, structures, and semantic relations from existing web resources. Web Mining enables semantic web vision, and the Semantic Web infrastructure improves web mining effectiveness. The effort behind the Semantic Web is to add semantic annotation to Web documents in order to access knowledge instead of unstructured material, allowing knowledge to be managed in an automatic way. Web Mining can help to learn definitions of structures for knowledge organization (e. g., ontologies) and to provide the population of such knowledge structures. The overall aim is thus not to replace the human, but rather to provide him with more and more support.

5.1. Exploiting Semantics for Web Mining

Semantic Web will change

- Content Mining: Gives a resounding view on contents and meaning of documents.
- Structure Mining: Provides a more relevant structure.
- Usage Mining: Imparts relevant information on actions of user.

The explicit encoding of semantics for mining the web in general improves intelligence of the systems.

5.2. Mining the Semantic Web

5.2.1. Semantic Web Content and Structure Mining

- Relational Data Mining (Inductive Logic Programming) involves looking for patterns that include multiple relations in a relational database.
- It comprises techniques for classification, regression, clustering, and association analysis.

5.2.2. Semantic Web Usage Mining

- Usage Mining can be enhanced further if the semantics are contained explicitly in the pages by referring to concepts of ontology.
- Log files can be mined, for instance to cluster users with similar interests in order to provide personalized views on the ontology.

VI. CONCLUSION

Web mining is the application of Data mining techniques to extract knowledge from the web resources. With the continued growth of the Web as an information source and as a medium for providing web services, Web Mining continues to play an ever expanding and inevitable role. Potentially lucrative & emerging area that holds particular promise is Semantic-Web Mining which is likely to make substantial gains in Web Mining research and practice.

REFERENCES

- [1] M. Kobayashi, and K. Takeda, .Information Retrieval on the Web., *ACM Computing Surveys*, Vol. 32, No.2, June 2000.
- [2] R. Kosala, and H. Blockeel, .Web Mining Research: A survey., *SIGKDD Explorations*, Vol. 2, Issue 1, July 2000, pp. 1-15.
- [3] B. Liu and K. Chang, Editorial: Special Issue on Web Content Mining, *SIGKDD Explorations*, Vol. 6, Issue2.
- [4] The Data Mining Encyclopedia, Idea Group Inc, 2006.
- [5] W. Bin, and L. Zhijing, .Web Mining Research., *Proc. 5th Int.l Conf. on Computational Intelligence and Multimedia Applications (ICCIMA.03)*, 2003.
- [6] Monika Henzinger and Steve Lawrence .Extracting knowledge from the World Wide Web., *Mapping Knowledge Domains*, 2003.
- [7] www.wikipedia.org., *Free-content encyclopedia on the Internet*.
- [8] S. Chakrabarti, .Data Mining for hypertext: A tutorial survey., *SIGKDD Explorations*, vol.1, no.2, 2000, pp.1-11.
- [9] S. Chakrabarti, .Mining the Web: Discovering knowledge from hypertext data., San Francisco, CA, 2003.
- [10] Monika Henzinger, .The Past, Present, and Future of Web Search Engines. *Proc. 31st International Colloquium, ICALP 2004*, Turku, Finland, July 12-16, 2004.
- [11] J. Srivastava, P. Desikan, and V. Kumar .Web Mining- Accomplishments and Future directions., *Proc. Nat.l Science Foundation Workshop on Next Generation Data Mining (NGDM.02)*, Baltimore, Maryland, 2002.
- [12] B. Mobasher, .Web Mining Overview., *Data Mining Encyclopedia*, Idea Group Inc, pp. 1206-1210.
- [13] www.w3.org, *The World Wide Web Consortium*.
- [14] T. Berners.Lee, J. Hendler, and O. Lassila, .The Semantic Web., *Scientific American*, vol.279, no.5, May 2001, pp.34-43.
- [15] B. Berendt, A.Hotho, and G.Stumme, .Towards Semantic Web Mining., *Proc. 1st Int.l Semantic Web Conf. (ISWC02)*, Sardinia, Italy, 2002.
- [16] A. M. Tjoa, A. Andjomshoaa, F. Shayeganfar, R. Wagner, .Semantic Web: Challenges and Requirements., *Proc. 16th Int.l Workshop on Database and Expert Systems Applications (DEXA.05)*, IEEE, 2005.