Semantic Web Mining and its application in Human Resource Management

Ridhika Malik[1], Kunjana Vasudev[2] and Udayan Ghose[3]

[1] University School of Information Technology, Guru Gobind Singh Indraprastha University
Delhi, Delhi, India
ridhikamalik@gmail.com

[2] Amity School of Business, Amity University
Noida, Uttar Pradesh, India
kunjananitk@gmail.com

[3] University School of Information Technology, Guru Gobind Singh Indraprastha University
Delhi, Delhi, India
g_udayan@lycos.com

Abstract

The Semantic Web is a project and vision of the World Wide Web Consortium to extend the current Web, so that information is given a well-defined meaning and structure, enhancing computers and people to work in cooperation. Semantic web mining is the combination of web mining and semantic web. The knowledge of semantic web makes web mining easier to achieve and can also improve the effectiveness of web mining. Semantic web mining technologies are being added to enterprise solutions to accommodate new techniques for discovering relationships across different database, business applications and Web services. Since this is an interdisciplinary concept in both engineering and management; we first review web mining, semantic web, semantic web mining and finally propose an application of semantic web mining in human resource management.
.

Keywords: *Semantic wining, ontology, RDF, engineering in management, human resource semantic web management system.*

1. Introduction

The Web has now become the tool for collaborative work and sharing of information throughout the web. Unfortunately the existing web poses a key challenge: the huge amount of data available is interpretable by humans only, the machine support is limited. It is highly desired that machines should be able to intervene and help while making a search on internet. The major obstacle in achieving this goal has been the fact that as such data available on internet is unstructured, it is disparate and is spread across the web in variant formats. From the human resource perspective of an organization, the company generally has a separate information systems for finance, resource management, learning and development and a separate one for business modeling and process management. Thus even if our database is robust enough and contains proper documentation of how to map columns of a database, yet the data interpreted can be interpreted only by humans and is not interpretable by a machines. Today it is almost impossible to retrieve information with a keyword search when the information is spread over several pages. To solve all such cases, semantic web has been proposed. Semantic Web [1] provides a common framework that allows data to be shared and reused across applications, enterprises and community boundaries. Metadata is "data about data". Metadata provides context for data and is used to facilitate the understanding, characteristics and management of data. In data processing, metadata is definitional data that provides information about data managed within an application or environment. When structured into a hierarchical arrangement, metadata is called an ontology or schema [5]. Both terms describe what is available in order to achieve our objectives. For instance, the arrangement of subject headings in a library catalog serves not only as a guide to finding books on a particular subject in the stacks, but also as a guide to what subjects are available in the library's own ontology and how more specialized topics are related to or derived from the more general subject headings. Metadata is frequently stored in a central location and is used to help organizations standardize their data.

Through this paper we aim to first review semantic web mining which is a combination of the semantic web and web mining. Web mining is [2]: Extract interested, useful patterns and implicit information from the WWW resources and behavior. The Semantic Web tries to make the data machine understandable, while Web Mining (semi-)automatically extracts the useful knowledge hidden in data, and makes it available as an aggregation of manageable proportions.

We next propose a Human Resource Semantic Web Management System (HRSWM) which enables the organizations to achieve their goals by transforming disparate, fragmented data into viable information capable of answering key questions. We define a system which combines the benefits of Resource Descriptive Framework (RDF) [3] with the high performance of database management systems and thus improves performance in retrieving information in real time. Data from across organizations is reorganized and recombined to report key information about information (such as: where is the useful information). For example in an organization with several departments, each department will provide RDF files to the HRSWM. When RDF file is introduced into the application, a corresponding set of tables will be created automatically. Afterwards, the information from these RDF files will be correlated, in order to create a summarized and standardized RDF file. Now queries can be launched in the standardized RDFs in SQL language. The resulting information architecture provides a unified view of the data sources in the organization. Thus, our HRSWM provides real-time analysis capabilities standardizing disparate organizational databases and increases "the speed to insight".

2. Web mining and semantic web

A. Web Mining

Web mining is the application of data mining techniques to the content, structure, and usage of Web resources. It is thus “the nontrivial process of identifying valid, previously unknown, and potentially useful patterns” [4] in the huge amount of Web data. 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 of invaluable help in the transformation from human-understandable content to machine-understandable semantics. Three areas of Web mining are commonly distinguished: web content mining, web structure mining and web usage mining [5], [6], and [7].

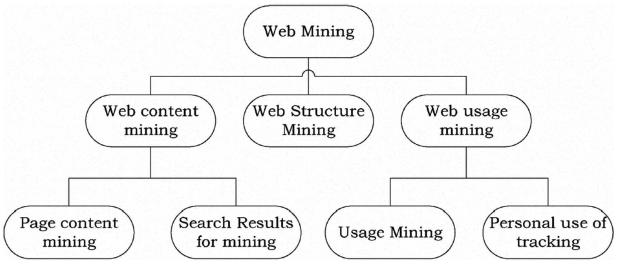

Figure 1: Web mining techniques

Web content mining is used to extract the text, image or other information and knowledge component of the web content. Which sites sell cars? Which pages are in Chinese? Which pages introduce the music, or introduce news? Search engines, intelligent agents and some recommend other applications 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. Content mining methods can be used for Ontology learning, mapping and merging ontologies and instance learning.

Web structure mining is used to extract the network topology information, that is, the link between pages of information. Web structure mining usually operates on the hyperlink structure of Web pages (for a survey). Mining focuses on sets of pages, ranging from a single Web site to the Web as a whole. Web structure mining exploits the additional information that is (often implicitly) contained in the structure of hypertext. Therefore, an important application area is the identification of the relative relevance of different pages that appear equally pertinent when analyzed with respect to their content in isolation. Which pages are linked to other pages? Which pages point to other page? Which collection of pages constitutes an independent entity? Such questions are answered by Web Structure Mining.
Web structure mining and Web content mining are often performed together, allowing exploiting simultaneously the content and the structure of hypertext.

Web usage mining is used to extract about the customer how to use the browser and use the page links. It extracts interesting patterns from the access to records of Web. For example, which pages are the client accesses? How long spent on each page? What next click on? What are the entry and exit routes? Each server retains the Web access log, recording information for the user access and interaction. Analysis of this data can help understand the user's behavior, thus improving the structure of the site, or to provide users with personalized services.

B. Semantic Web

The Semantic Web is based on a vision of Tim Berners-Lee, the inventor of the WWW. Berners-Lee suggests semantic web to enrich the Web by machine-understandable 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-understandable information can point the search engine to the relevant pages and can thus improve both precision and recall. The following steps show the direction where semantic Web is heading:

1. Providing a common syntax for machine understandable statements.
2. Establishing common vocabularies.
3. Agreeing on a logical language.
4. Using the language for exchanging proofs.

Berners-Lee suggested a layer structure for the Semantic Web. This structure reflects the steps listed above. It follows the understanding that the semantic Web can be realized in an incremental fashion and each step will provide added value.

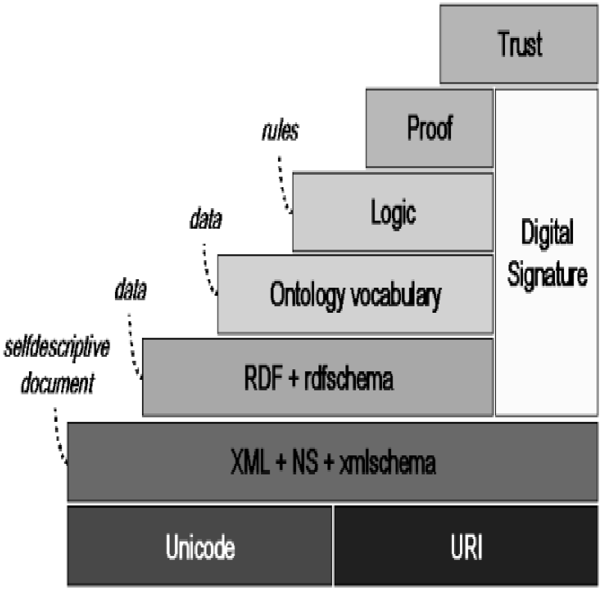

Figure 2: Semantic Web Layers

According to Berners-Lee's vision, the semantic network Constituted by seven levels is a layered architecture. The first layer of URI and Unicode is the basis for the structure of the entire system. Unicode is responsible for processing resources encoding, URI is responsible for resource identification, which allows precise retrieval of information possible. The Second layer of XML + NS (Namespace) + XML Schema, is responsible for representing the content and structure of data from the linguistic to separate the performance format, the data structure and content of the network information form through the use of a standard format language. The third layer of RDF + RDF Schema, which provides a semantic model used to describe the information on the Web and type. The fourth layer of ontology vocabulary layer is responsible for the definition of shared knowledge and describes the semantic relationships between the various kinds of information to reveal the semantic between information itself and information. The fifth layer of logic layer is responsible for providing axioms and inference principles to provide the basis for intelligent services. The sixth layer of proof and the seventh layer of trust are responsible for providing authentication and trust mechanisms. Digital signatures and encryption technology used to detect changes in the document situation is a means to enhance Web security.

Semantic Web is known as Web3.0, it is based on RDF to integrate variety of applications of XML-syntax. It uses uniform resource identifier as a naming mechanism. Semantic Web is just an extension of the current Web and is not a new Web. The research focus is how the information can only be changed from the form that a computer can read to the form that a computer can understand and deal with, that is with the semantics, so that the computer and people can work together. Web resources (such as Web pages, Web service) for the use of ontology annotation terms are important prerequisites for goal to achieve the semantic Web. Ontology-based semantic annotation using ontology defined by experts support the content creator to add semantic metadata in the Web page, so content can be understood by people and machines, as compared with the general public, this is a marked top-down classification. Semantic Web which can be seen as a new generation of information infrastructure is a new distributed intelligent network platform based on semantic information processing.

3. Semantic Web Mining

Semantic Mining [8] is a series of semantic analysis of information resources and users' question by advanced intelligence theory and technology, through mining its deep semantics, in order to fully and accurately express knowledge resources and user needs. The mined data can be used in various distributed, heterogeneous databases, data warehouses, knowledge Bases to search and retrieve information in intelligent processing to return the most relevant results of the semantic retrieval mechanism.

Semantic-based Web data mining combines semantics that are extracted from existing Web data extraction or uses existing semantic structures with Web Mining. Web mining results help to build the semantic Web, the Semantic Web mining knowledge makes it easier to achieve and improves the effectiveness of Web mining. Parallel to the Web mining, semantic-based Web Mining we can be divided into semantic Web content mining, Semantic Web structure mining and semantic Web usage mining categories:-

- *Semantic Web content and structure mining*- In the web mining based semantic network, the differences between content mining and structure mining are almost vanished, so we refer to them collectively as the semantic Web content and structure mining. Thus, the traditional data mining can easily be transferred to the Semantic Web content and structure mining. This is achieved through-
  a) Ontology learning
  b) Mapping and merging ontologies
  c) Instance learning
  d) Semantics created by structure

- *Semantic Web usage mining*- In the Semantic Web environment, we can give clear semantics to user behavior. On this basis, we can find the users with the same interest, which provides users with ontology-based personalized view to improve the Web usage mining results

4. Application of Semantic Web mining in Human Resource Management- Human Resource Semantic Web Management System (HRSWM)

The main goal of this concept is to integrate the Semantic Web with the actual needs of a corporate process by combining the strong points of semantic technologies with the high performance of database management systems. In our Humana Resource Semantic Web Management we consider an enterprise consisting of several departments. Each department will provide RDF files to the HRSWM.When an RDF file is introduced into the application, a corresponding set of tables will be created automatically. An information model (ontology) will be created based on data schema provided by different departments of the enterprise. Variant, unstructured and disparate data from several departments is gathered from RDF files of each department and uniform information models relating all departments are prepared. The end-users can then query this semantic (metadata) model, which comprise any number of RDF files or ontologies. The resulting information architecture provides a unified view of the data sources in the organization.

The architecture of HRSWM consists of three main layers:

- The *Consumer Layer*, which interfaces the resources (the RDF source files);
- The *Middle Layer*, which holds certain resources. This layer is the mediator which enables a common communication language between the source of information and the HRSWM, and
- The *Warehousing Component*, which stores and prepares the information for future browsing or updates.

In our architecture, the Warehousing Component and the Consumer Layer is permanently connected and certain jobs are executed on a scheduled calendar to ensure an automated process.

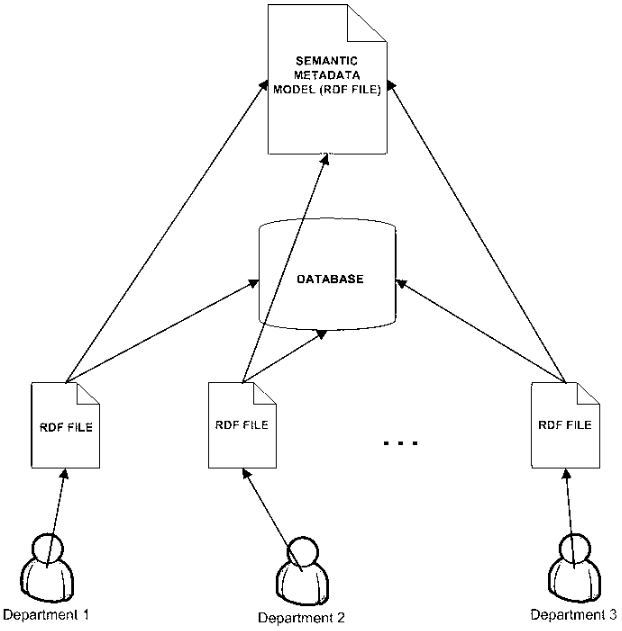

Figure 3: Human Resource Semantic Web Management

4. Conclusion

The need to search and derive greater business knowledge from existing database repositories and applications has become a high priority in most of the organizations.

Semantic technologies are being added to enterprise solutions to accommodate new techniques for discovering relationships across different database, business applications and Web services.

The HRSWM system proposed in this paper has the following advantages: access semantic data through SQL and do mixed queries – relational and RDF queries in the same SQL statement, improved data access, scalability and platform independence.

References

[1] Berners-Lee, T., Hendler, J., and Lassila, O., The Semantic Web, Scientific, American, USA, 2001.

[2] Wen-Wei Chen, "Data Warehouse and Data Mining Tutorial",[M], Beijing: Tsinghai University Press, 2008.A. Name, and B. Name, "Journal Paper Title", Journal Name, Vol. X, No. X, Year, pp. xxx-xxx.

[3] RDF, http://www.w3.org/RDF/, 01.06.2009

[4] Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P., From data mining to knowledge discovery. In Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P, editors, Advances in Knowledge Discovery and Data Mining, pages 1–34. AAAI / MIT Press, Cambridge, MA, 1996.

[5] Osmar, R., from resource discovery to knowledge discovery on the internet. Technical Report TR 1998-13, Simon Fraser University, 1998

[6] Kosala, R. and Blockeel, H.,. Web mining research: A survey. SIGKDD Explorations, 2(1), 2000.

[7] Srivastava, J., Cooley, R., Deshpande, M., Web usage mining: discovery and application of usage patterns from web data. SIGKDD Explorations, 1(2):12–23, 2000.

[8] Zhang Hui, ed, "Ontology-based Semantic Web Mining Technology."[D], computer development and applications, 2009,