Integrating Semantic Web and Web Mining into Semantic Web Mining

Lisana BERBERI

Department of Informatics and Mathematics, University of Shkoder, "LUIGJ GURAKUQI", Shkoder, Albania
Email: lisanaberberi@yahoo.com – Phone: +355(673153563)

ABSTRACT

We know semantic web makes information more readable and meaningful to people by making it more understandable to machines and web mining is the application of data mining techniques to discover patterns from the Web.

We are aware websites on the Internet are increasing, daily, in size and complexity, which makes rather difficult that specific information, can be easily found.

In this paper will be described how we can integrate semantic web and web mining into semantic web mining.

Keywords: *Semantic Web, Web Mining*

INTRODUCTION

The Internet created a standard way for computers to communicate with one another, in other words it gave a voice to computers so they may talk to each other and exchange information.

If a computer understands the *semantics* of a document, it doesn't just interpret the series of characters that make up that document: it understands the document's *meaning.*

The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. Semantic Web can be described as an efficient way to represent data on the World Wide Web, or as a database that is globally linked, in a manner understandable by machines, to the content of documents on the Web.

So the Semantic Web holds a great deal in making our life easier by helping computers help us to get what we want.

Currently the Semantic Web, which was conceived by Tim Berners-Lee, the inventor of the World Wide Web, is the focus of a W3C working group.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining is mining of data related to World Wide Web (WWW); the data are presented in web pages or related to web activity, the problems which the users face in the website are:

- Detection of relevant information.
- Discover of existing but 'hidden' knowledge.

When the customers log in to the website, they want to obtain information from the web. So, to solve this management problem we need the right techniques and methods that derive from different areas such as: Expert Systems (ES), Artificial Intelligent (AI), Database (DB), and one of the Information Retrieval (IR) methods such as Structure Query Language (SQL). In brief, Web Mining: automated discovery and analysis of useful information from web documents and services using one of the data mining techniques.

SEMANTIC WEB AND WEB MINING TECHNOLOGIES

SEMANTIC WEB

Semantic Web technologies help separate meanings from data, document content, or application code, using technologies based on open standards.

Semantic technologies represent meaning using *ontologies* and provide reasoning through the relationships, rules, logic, and conditions represented in those ontologies.

To represent the Semantic Web, we'll use the following technologies:

- A global naming scheme (URIs)
- A standard syntax for describing data (RDF)
- A standard means of describing the properties of that data (RDF Schema)
- A standard means of describing relationships between data items (ontologies defined with the OWL Web Ontology Language)

A global naming scheme: URIs

A *URI* is simply a Web identifier, like the strings starting with http or ftp that we often see on the World Wide Web. Anyone can create a URI, and the ownership of URIs is clearly delegated, so they form an ideal base technology on top of which to build a global Web. In fact, the World Wide Web is such a thing: anything that has a URI is considered to be "on the Web." Every data object and every data schema/model in the Semantic Web must have a unique URI.

A Uniform Resource Locator (URL) is a URI that, in addition to identifying a resource, provides a means of acting upon or obtaining a representation of that resource by describing its primary access mechanism or network location.

A standard syntax to describe data: RDF

RDF is a specification that defines a model for representing the world, and syntax for serializing and exchanging that model. The W3C has developed an XML serialization for RDF. RDF XML is the standard interchange format for RDF on the Semantic Web, although it is not the only format.

RDF provides a consistent, standardized way to describe and query Internet resources, from text pages and graphics to audio files and video clips. It offers syntactic interoperability, and provides the base layer for building a Semantic Web. RDF defines a directed graph of relationships.

WEB MINING

Three areas of Web mining are commonly distinguished: content mining, structure mining, and usage mining.

Content mining: *Web content mining* is a form of text mining. The primary Web resource that is being mined is an individual page. Web content mining can take advantage of the semi-structured nature of Web page text. The HTML tags of today's Web pages, and even more so the XML markup of tomorrow's Web pages, bear information that concerns not only layout, but also logical structure. Web content mining can be used to detect co-occurrences of terms in texts.

Structure mining: *Web structure mining* usually operates on the hyperlink structure of Web pages. The primary Web resource that is being mined is a set 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 *hyper*text.

Usage mining: In *Web usage mining*, the primary Web resource that is being mined is a record of the requests made by visitors to a Web site, most often collected in a Web server log. The content and structure of Web pages, and in particular those of one Web site, reflect the intentions of those who have authored and designed those pages, and their underlying information architecture. The actual behavior of those who use these resources may reveal additional structure.

It is useful to combine Web usage mining with content and structure analysis in order to "make sense" of observed frequent paths and the pages on these paths. This can be done using a variety of methods. Some methods classify pages in terms of a pre-defined ontology, while others rely on the extraction of keywords found in these pages, and subsequent human naming of the keyword clusters represented by frequent paths.

Table 1. Web Mining Categories

	Web Mining			
	Web Content Mining		**Web Structure Mining**	**Web Usage Mining**
	Information Retrieval View	**Database View**		
View of Data	-Unstructured -Semi-structured	-Semi structured -Web site as DB	-Links structure	-Interactivity
Main Data	-Text Documents -Hypertext Documents	-Hypertext documents	-Links structure	-Server logs -Browser logs
Representation	-Bag of words, n-grams -Terms, phrases -Concepts or ontology -Relational	-Edge-labeled graph (OEM) -Relational	-Graph	-Relational table -Graph
Method	-TFIDF and variants -Machine learning -Statistical (including NLP)	-Proprietary algorithms -ILP -(Modified) association rules	-Proprietary algorithms	-Machine learning - Statistical - (Modified) association rules
Application Categories	-Categorization -Clustering -Finding extraction rules	- Finding frequent sub-structures -Web site schema discovery	-Categorization -Clustering	-Site construction -Site adaptation

	-Finding patterns in text -User modeling			-Site management -Marketing -User modeling

SEMANTIC WEB MINING

Semantic web mining essentially is mining the information pertaining to the semantic web. This means mining Web pages so that machine can better understand the information. It also means mining the data sources to develop an effective semantic Web.

Figure 1 illustrates semantic web mining. It shows mining various XML and RDF documents as well as mining ontologies and metadata.

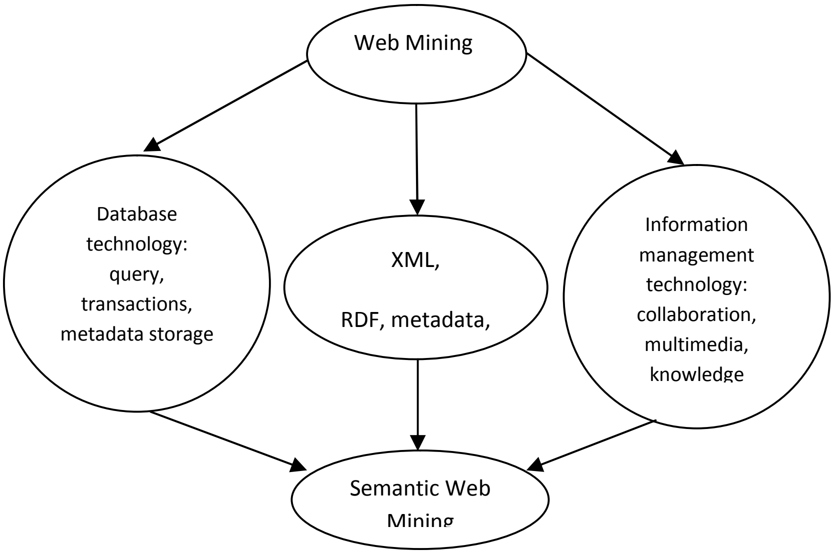
Web Mining
Database technology: query, transactions, metadata storage
XML,
RDF, metadata,
Information management technology: collaboration, multimedia, knowledge
Semantic Web Mining

Figure 1 Semantic Web Mining Technologies

To give a more precisely idea about the combination of these two areas, Semantic Web and Web Mining, we can view an example taken from [Berendt1, Hotho, Stumme, D–10178 Berlin, Germany] which provides ontology-based access to tourism Web pages.

We will first learn an ontology using Web Mining, then fill the ontology with instances by again using Web Mining, and finally mine the resulting data in order to gain further insights. One may split the first step, ontology learning, in two sub-steps. First a concept hierarchy is established using the knowledge acquisition method OntEx (Ontology Exploration). OntEx takes as input a set of concepts, and provides as output a hierarchy on them. This output is then the input to the second sub-step, together with a set of Web pages. In Fig. 3 is described how association rules are mined from this input, which leads to the generation of relations between the ontology concepts.

The association rules are used to discover combinations of concepts which frequently occur together. These combinations hint at the existence of conceptual relations. They are suggested to the user.

In the example shown in the figure, automatic analysis has shown that three concepts frequently co-occur with the concept "area". Since the ontology bears the information that the concept "wellness hotel" is a sub concept of the concept "hotel", which in turn is a sub concept of "accommodation", the inference engine can derive that only one conceptual relation needs to be inferred based on these co-occurrences: the one between "accommodation" and "area".

Human input is then needed to identify that an accommodation "hasLocation" that is an area, i. e., to specify a name for the generalized conceptual relation.

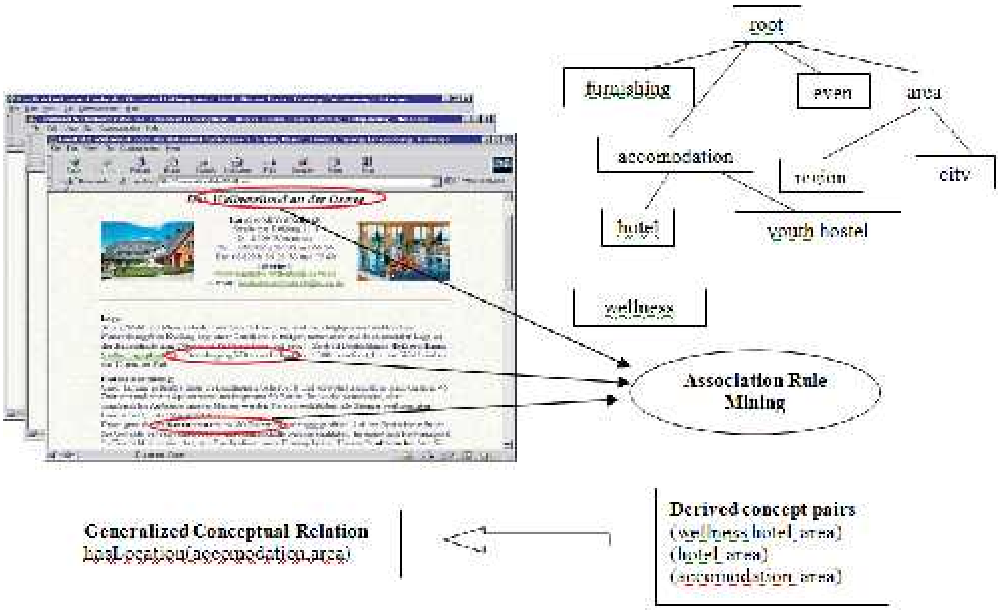

Figure 3. Step 1: Mining the Web for learning ontologies.

In the second step, *the ontology is filled*. In this step, instances are extracted from the Web pages, and the relations from the ontology are established between them using techniques (see Fig. 4). Beside the ontology, the approach needs tagged training data as input. Given this input, the system learns to extract instances and relations from other Web pages and from hyperlinks. After the second step, we have an ontology and a knowledge base, i. e., instances of the ontology concepts and relations between them.

These data are now input to the third step, in which *the knowledge base is mined*. Depending on the purpose, different techniques maybe applied. One can for instance derive relational association rules, as described in detail Fig. 5.

In the example shown in Figure 5, a combination of knowledge about instances like the Wellnesshotel and its SeaView golf course, with other knowledge derived from the Web pages' texts, produces the rule that hotels with golf courses often have 5 stars.

More precisely, this holds for 89% of hotels with golf courses, and 0.4% of all hotels in the knowledge base are five star hotels owning a golf course.

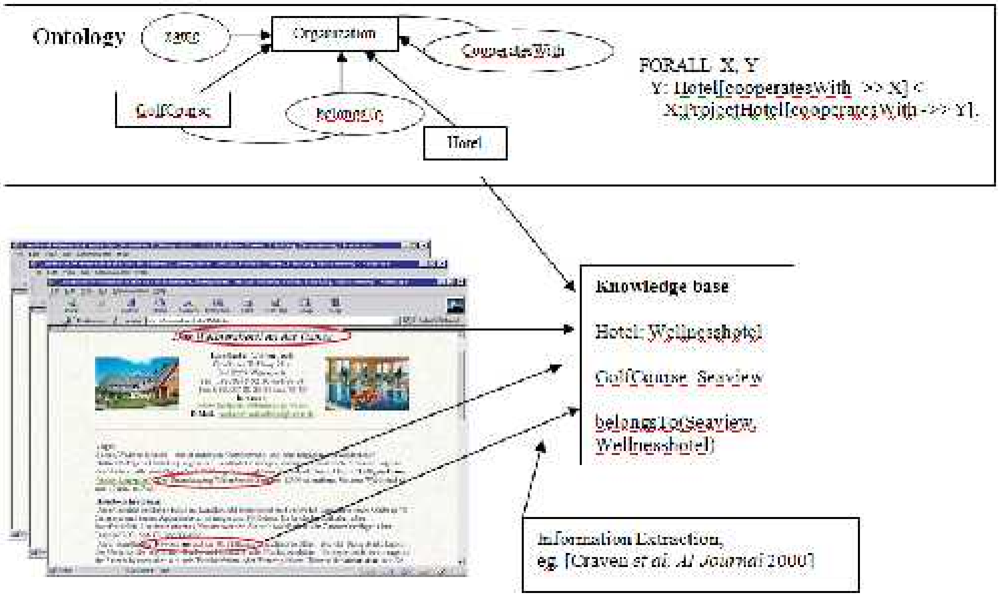

Figure. 4. Step 2: Mining the web for filling the ontology.

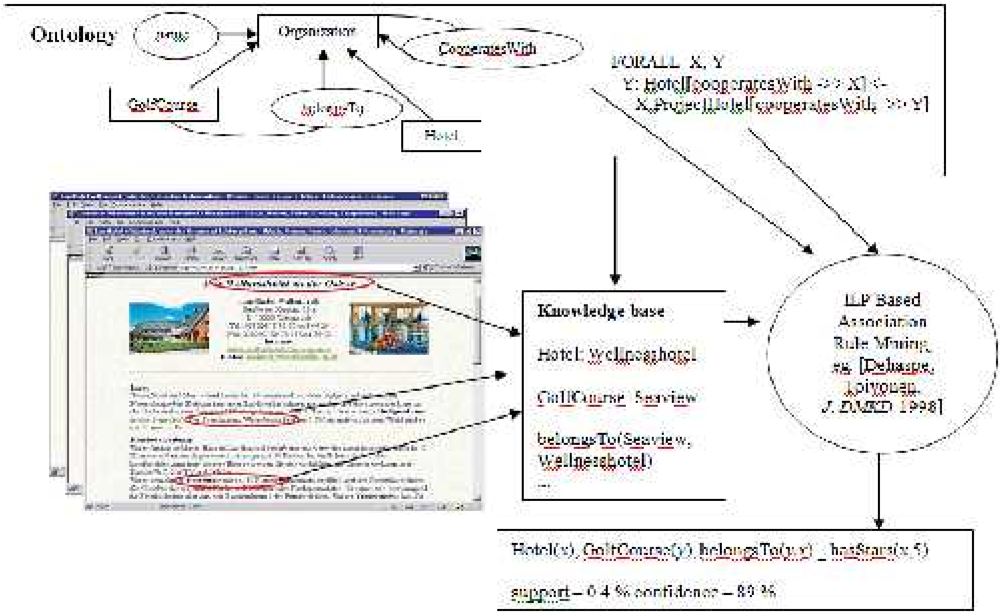
Ontology
name
Organization
CooperatesWith
GolfCourse
belongsTo
Hotel
FORALL X, Y
Y: Hotel[cooperatesWith ->> X] <-
X:ProjectHotel[cooperatesWith ->> Y]
Knowledge base
Hotel: Wellnesshotel
GolfCourse: Seaview
belongsTo(Seaview, Wellnesshotel)
...
ILP Based Association Rule Mining,
eg. [Dehaspe, Toivonen, J. DMKD 1998]
Hotel(x), GolfCourse(y), belongsTo(y,x) _ hasStars(x,5)
support – 0.4 % confidence – 89 %

Figure 5. Step 3: Using the ontology for mining again.

CONCLUSIONS

The main idea is to improve the results of Web Mining by exploiting the new semantic structures in the Web from one side; and to make use of Web Mining for building up the semantic Web, from the other side.

From a business perspective, one of the major goals of web usage mining is the personalization to individual users on a massive scale, often known as mass customization.

By customizing the page to the user's profile, not only does the user get what he/she wants, but marketing and sales advertisements are directed at the correct demographic and thus generate maximum return.

These techniques are important to business which relies on e-commerce, so Web Usage Mining can design cross marketing strategies across products, evaluate promotional campaigns, predict user behavior based on previously learned rules and users 'profile and present dynamic information to users based on their interests and profiles.

In recent years, there have been build various models of e-learning platform with learning resources recommendation based on web usage mining for mining the server logs and the database to find the users' usage patterns, to provide users with more personalized services.

REFERENCES

[1] Bettina Berendt1, Andreas Hotho, and Gerd Stumme, Towards Semantic Web Mining, Institute of Information Systems, Humboldt University Berlin Spandauer Str. 1, D–10178 Berlin, Germany

[2] L. Dehaspe and H. Toivonen. Discovery of frequent datalog patterns. Data Mining and Knowledge Discovery, 3(1):7–36, 1999.

[3] S. Brin, L. Page, "The anatomy of a large-scale hyper-textual Web search engine". In the 7th International World Wide Web Conference, Brisbane, Australia, 1998.

[4] E. Colet, "Using Data Mining to Detect Fraud in Auctions", DSStar, 2002.Chow

[5] R. Cooley, B. Mobasher, J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", in Proceedings of the 9th IEEE International Conference on Tools With Artificial Intelligence (ICTAI '97), Newport Beach, CA, 1997.

[6] Naveen Balani, Technical Architect, Webify Solutions, Ontologies form the backbone of a whole new way to understand online data

[7] Michael Kernahan, Different Strategies for Web Mining, London

[8] Xinjin Li; Sujing Zhang; Coll. of Teacher Educ., Zhejiang Normal Univ., Jinhua, China, Application of Web Usage Mining in e-learning Platform

[9] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD thesis, University of Minnesota, Faculty of the Graduate School, 2000.

[10] G. Stumme, R. Taouil, Y. Bastide, N. Pasqier, and L. Lakhal. Computing iceberg concept lattices with titanic. J. on Knowledge and Data Engineering (in print), 2002.