Web mining accomplishments and the semantic web potentials

Amani Anwar Saad
Computer and Systems Engg., Dept., Faculty of Engg., Alexandria University, Alexandria, Egypt
Amanisaad2008@gmail.com

Web mining research is actually a converging research area from several research communities, such as Database (DB), Information Retrieval (IR) and Artificial Intelligence (AI) especially from machine learning and Natural Language Processing (NLP). Web Mining aims at extracting useful knowledge from the web in order to solve the information overload problem facing the web users. The Semantic Web has evolved as the second-generation WWW, enriched by machine process-able information which supports the user in his tasks. Thus Semantic Web mining is an evolving area which combines research in both areas of Semantic Web as well as Web mining in the first generation WWW. In this paper it is argued that Semantic Web Mining should be added to the taxonomy of Web mining categories introduced in the previous literature. We present a survey for the Web mining field, techniques, applications, accomplishments and potentials revisited after the emergence of the Semantic Web.

التنقيب عن المعرفة في الشبكة العنكبوتية العالمية WWW هو مجال من مجالات البحث المنبثقة عن العديد من المجتمعات البحثية مثل قواعد البيانات، استرجاع المعلومات، الذكاء الاصطناعي، معالجة اللغات الطبيعية وطرق تطوير ذكاء اللألة. وهو مجال بحث يهدف الي استخدام افضل للشبكة العنكبوتية وايجاد حلول لمشاكل مستخدمي الشبكة من تضخم المعلومات وتعدد المصادر وعدم القدرة علي التحقق من مصداقيتها وملاءمتها للهدف من البحث. ولقد تم تطوير الشبكة الدلالية Semantic Webوهي تمثل الجيل الثاني من الشبكة العنكبوتية WWW من اجل نفس الأهداف وذلك عن طريق اضافة بعض المعلومات التي تستطيع ان تتعامل معها وتعالجها اللألة وبالتالي تستطيع تقديم العون والارشاد للمستخدم user وذلك عن طريق اضافة الكثير من الدلالات والأوصاف للبيانات الموجودة علي الشبكة مثل استخدام لغة XML بدلا من HTML وما شابه ذلك. ومن الجدير بالذكر أن هذين المجالين للبحث قد تم اندماجهما الي حد كبير في السنوات الأخيرة بحيث اصبح كل منهما يخدم اللأخر كما يستفيد مما تحقق من تقدم فيه. ومن أهداف هذا البحث مناقشة وعرض ما تم انجازه في كل من المجالين واقتراح تعديل التصنيف الحالي لطرق التنقيب عن المعرفة في الشبكة العنكبوتية العالمية و ذلك بناءا علي ما تم انجازه في مجال تطوير الشبكة الدلالية Semantic Web. ومن ثم اقتراح العديد من نقاط البحث المستقبلية في هذا المجال الغني.

Keywords: Web mining, Information retrieval, Ontologies, Machine intelligence, Semantic web, Software agents.

1. Introduction

The World Wide Web (WWW) is a popular and interactive medium to disseminate information today. It is a huge, widely distributed, highly heterogeneous, semi-structured, interconnected, evolving, hypertext/hypermedia information repository.

With the huge amount of information available online, the Web is a fertile area for data mining research. The spectacular ascent in the size and popularity of the Web has subjected traditional information retrieval techniques to an intense stress–test. The Web contains millions of HTML pages, Extensible Markup Language (XML) pages, as well as different multimedia data. In the year 2000, there was six terabytes of data on about three million servers [1]. Almost one million pages are added daily, a typical page changes in a few months, and several hundred gigabytes change every month. Even the largest search engines could index a small portion of the accessible Web.

Information users are drowning in information and facing an information overload problem. Faced with the ever-growing importance of the Web, users need a “better Web” that meets their expectations. Users expect intelligent processing such as search engines that recognize their true information needs and a broad and accurate coverage of all aspects of their lives. Thus, users need better processing and analysis tools for the Web content as well as better modeling or representation tools for real world objects that are described in the Web Space such as jobs, skills, places, diseases, etc.

As a consequence to these needs, two major results have occurred: the development of research in Web mining techniques as well as the evolution of the Semantic Web. Indeed, Web mining techniques fulfill the first need and provide the processing and analysis tools that could be used to solve the information overload problem. In a direct way, web mining techniques can solve problems like deciding whether a piece of news is relevant to a user subscribing at a certain News group or not based on the analysis of his profile. Moreover, Web mining techniques can be used indirectly as taking part in bigger applications that address some problems like creating index terms for the Web search services, finding relevant information for a user, creating new knowledge from the retrieved information, personalization of information and providing assistance to the Web site designer to model the user behavior so as to target better usability parameters of his site and achieve his profit targets like in e-commerce sites.

On the other hand, to fulfill the second need for users, the Semantic Web has evolved during the last decade as the second-generation WWW, enriched by machine process-able information which supports the user in his tasks [2]. However, given the enormous size even of today's Web, it is impossible to manually enrich all of these resources. Thus, automated schemes for learning the relevant information are increasingly being used.

Web Mining aims at discovering insights about the meaning of Web resources and their usage. Given the primarily syntactical nature of the data being mined, the discovery of meaning is impossible based on these data only. Therefore, formalizations of the semantics of Web sites and navigation behavior are becoming more and more common.

As a further result, Semantic Web mining is an evolving area which combines research in both areas of Semantic Web as well as Web mining in the first generation WWW. However, Semantic Web Mining has two broad fields in which semantics help Web mining techniques or Web mining techniques help in extracting Semantics.

In this paper, we argue that Semantic Web Mining should be added to the taxonomy of Web mining categories introduced in previous literature. We present a survey of the Web mining field, techniques, applications, accomplishments and potentials revisited after the emergence of the Semantic Web.

The rest of this paper is organized as follows: Section 2 presents the Semantic Web, Section 3 presents an overview of the Web Mining process, its subtasks and elaborates upon the pattern discovery process which is the core of any knowledge discovery process. Section 4 suggests a new taxonomy for Web mining categories. Section 5 presents Web Content Mining and specially Multimedia data mining, and illustrates the problem of querying the Web for resources and knowledge. Section 6 presents Web Structure Mining, then. Section 7 presents a combined approach for mining content as well as structure which deals with the problem of Hypertext mining. Section 8 presents Web Usage Mining. Section 9 presents Semantic Web Mining and its new potentials. Afterwards, the close relationship between Web Mining and the Software Agent Paradigm is illustrated in Section 10. Finally, Section 11 summarizes this paper and presents future research directions in this fertile area.

2. The semantic web

The Semantic Web is the second-generation WWW, enriched by machine-processable information which supports the user in his tasks. It is based on a vision of Tim Berners-Lee, the inventor of the WWW. The great success of the current WWW leads to a new challenge: A huge amount of data is interpretable by humans only; machine support is limited. Berners-Lee suggests to enrich the Web by machine process-able information which supports the user in his tasks. Indeed, today's search engines are already quite powerful, but still too often return excessively large or inadequate lists of hits. Machine process-able information can point the search engine to the relevant pages and can thus improve both precision and recall. For instance, today it is almost impossible to retrieve information with a keyword search when the information is spread over several pages. Consider, e.g., the query for Web Mining experts in a company intranet, where the only explicit information stored is the relationship between people and the courses they attended on one hand, and between courses and the topics they cover on the other hand. In that case, the use of a rule stating that people who attended a course which was about a certain topic have knowledge about that topic might improve the results [2].

The process of building the Semantic Web is currently an area of high activity. Its structure has to be defined, and this structure then has to be filled with life. In order to make this task feasible, one should start with the simpler tasks first. The following steps show the direction where the Semantic Web is heading:

- Providing a common syntax for machine understandable statements.
- Establishing common vocabularies.
- Agreeing on a logical language.
- Using the language for exchanging proofs.

Berners-Lee suggested a layered structure for the Semantic Web. Fig. 1 illustrates these layers [2]. This structure reflects the steps listed above. It follows the understanding that each step alone will already provide added value, so that the Semantic Web can be realized in an incremental fashion. This architecture is discussed in detail in [3] which also addresses research questions. On the first two layers, a common syntax is provided. Uniform Resource Identifiers (URIs) provide a standard way to refer to entities, while Unicode is a standard code for exchanging symbols. The XML fixes a notation for describing labeled trees, and XML Schema allows the definition of grammars for valid XML documents. XML documents can refer to different namespaces to make explicit the context (and therefore meaning) of different tags. The formalizations on these two layers are nowadays widely accepted, and the number of XML documents is increasing rapidly. While XML is one step in the right direction, it only formalizes the structure of a document and not its content. The Resource Description Framework (RDF) can be seen as the first layer where information becomes machine-understandable: According to the W3C recommendation, RDF "is a foundation for processing metadata; it provides interoperability between applications that exchange machine- understandable information on the Web".

RDF documents consist of three types of entities: Resources, properties, and statements. Resources may be Web pages, parts or collections of Web pages, or any real-world objects, which are not directly part of the WWW. In RDF, resources are always addressed by URIs. Properties are specific attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource form an RDF statement. A value is either a literal, a resource, or another statement. Statements can thus be considered as object-attribute-value triples. The data model underlying RDF is basically a directed labeled graph. RDF Schema defines a simple modeling language on top of RDF which includes classes, is-a relationships between classes and between properties, and domain/range restrictions for properties. RDF and RDF Schema are written in XML syntax, but they do not employ the tree semantics of XML.

XML and XML schema were designed to describe the structure of text documents, like HTML, Word, Star-Office, or LATEX documents. It is possible to define tags in XML to carry metadata but these tags do not have formally defined semantics and thus their meaning will not be well-defined. It is also difficult to convert one XML document to another one without any additionally specified semantics of the used tags. The purpose of XML is to group the objects of content, but not to describe the content. Thus, XML helps the organization of documents by providing a formal syntax.

The next layer is the ontology vocabulary. Following [4], an *Ontology* is *"an explicit formalization of a shared understanding of a conceptualization".* This high-level definition is

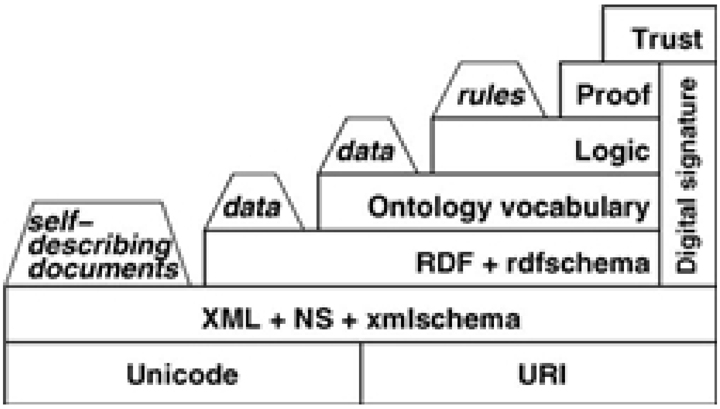

Fig. 1. The semantic web layers.

realized differently by different research communities. However, most of them have a certain understanding in common, as most of them include a set of concepts, a hierarchy on them, and relations between concepts. Most of them also include axioms in some specific logic. The Karlsruhe Ontology framework KAON. is built in a modular way, so that different needs can be fulfilled by combining parts.

Logic is the next layer according to Berners-Lee. Today, most research treats the ontology and the logic levels in an integrated fashion because most ontologies allow for logical axioms. By applying logical deduction, one can infer new knowledge from the information which is stated explicitly.

Proof and trust are the remaining layers. They follow the understanding that it is important to be able to check the validity of statements made in the (Semantic) Web, and that trust in the Semantic Web and the way it processes information will increase in the presence of statements thus validated. Therefore, the author must provide a proof which should be verifiable by a machine. At this level, it is not required that the machine of the reader finds the proof itself, it ‘just’ has to check the proof provided by the author.

3. Web mining

Web Mining is the overall process of discovering potentially useful and previously unknown information or knowledge from the Web data. It is the use of data mining techniques to automatically discover and extract information from the Web documents and services [5].

Similar to [5] Web Mining can be decomposed into the following subtasks:

- *Resource finding:* the task of retrieving intended Web documents. It is the process of retrieving data that is either online or offline from the text sources available on the Web such as newsletters, newsgroups, the text content of HTML documents by removing HTML tags, and also the manual selection of Web resources. This is a typical Information Retrieval process.
- *Information selection and pre-processing:* automatically selecting and pre-processing specific information from retrieved Web resources. This is any kind of transformation process that is done on the original data obtained in the IR process. For example, it could be a pre-processing step such as removing stop words or stemming or it might be aiming at obtaining a desired representation such as changing the representation to relational or finding phrases in the training corpus.
- *Generalization (Pattern discovery):* automatically discovers general patterns at individual Web sites as well as across multiple sites. This step is when actually machine learning and data mining techniques are used for pattern discovery.
- *Pattern analysis*: validation and /or interpretation of the mined patterns.

3.1. Pattern discovery

The Pattern discovery process draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. This section describes the kinds of mining activities that have been applied to the Web domain. For a survey of algorithms and techniques derived from these fields and used for knowledge discovery in databases see [6].

The most important techniques used in Web mining are:

- *Statistical Analysis:* Statistical techniques are the most common method to extract knowledge about visitors of a Web site. By analyzing the session file, one can perform different kinds of descriptive statistical analysis (frequency, mean, median,…etc) on variables such as page views, viewing time and length of a navigational path. Many Web traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages of a Web site or the average length of a path through a site.
- *Association Rules:* Association rule generation can be used to relate pages that are most often referenced together in a single server session. In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding a specified threshold. These pages may not be related via hyperlinks. For example, using the Apriori algorithm [7] correlation between users who visited a page containing electronic products to those who access a page about sporting equipment may be revealed. Aside from being applicable for business and e-commerce, these rules may help Web designers to restructure their Web sites.
- *Clustering:* Clustering is a technique to group together a set of items having similar characteristics (sometimes called unsupervised learning). In Web Usage domain, there are two kinds of interesting clusters to be discovered: *usage clusters* and *page clusters*. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. This is useful in order to perform market segmentation in e-commerce applications, and also to provide personalized Web content to users. On the other hand, clustering of pages discovers groups of pages having related content. This knowledge is useful for internet search engines and web assistance providers or recommender systems. Clustering may also be used for adaptive web sites as in [8].
- *Classification:* Classification is the task of mapping a data item into one of several predefined classes (sometimes called Supervised learning). In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires the extraction and selection of features that best describe the properties of a given class or category. Classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve-Bayes classifier, and k-nearest neighbor classifiers [9]. For example, classification on Web server logs may lead to discover the following rule: 30% of users who placed an on-line order in Music products are in the 18-25 age group and live in North America.
- *Sequential Patterns:* The technique of sequential patterns discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. Other types of temporal analysis on sequential patterns include trend analysis, similarity analysis, and change point detection. In [10] pruning strategies for sequential pattern analysis are discussed to decrease the total cost of mining.
- *Dependency Modeling:* Dependency modeling is another useful pattern discovery task in web mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the Web domain. For example, modeling the different stages a visitor undergoes while shopping in an online store based on the actions chosen (i.e., from a casual visitor to a serious potential buyer). There are several probabilistic learning techniques that can be employed to model the browsing behavior of users. Such techniques include Hidden Markov Models and Bayesian Belief Networks

[9]. Modeling of Web usage patterns not only provides a theoretical framework for analyzing the behavior of users but is potentially useful for predicting future Web resource consumption. Such information may help developing strategies to increase the sales of products offered by a Web site or improve the navigational convenience of users.

4. Web mining taxonomy

According to [11], Web data can be classified into three major categories which are: content, structure and usage. *Content* data is the real data in the web pages, i.e., the data the web page is designed to convey to the users. This not only consists of text and graphics but it may also include, image, audio clips, video clips, etc.. *Structure* is the data which describes the organization of the content. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. The inter-page structure also describes how the web pages in one web site or multiple sites are linked together. *Usage* data describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of access. In [11], a *User Profile* is defined as the data that provides demographic information about users of a Web site. This includes registration data and customer profile information. This is especially important for Web Usage mining.

The work done in [12-14], categorizes web mining into three areas of interest, according to which part of the Web to mine: Web Content Mining, Web Structure Mining, and Web Usage Mining (WUM).

4.1. Web Content Mining (WCM)

WCM describes the discovery of useful information from the Web contents/data/documents. However, what consists Web content could encompass a very broad range of data. All types of services previously known such as Gopher, FTP and Use-nets are now accessible from the Web. Moreover, Digital Libraries, and many companies are transforming their business and services electronically. Thus, many legacy systems are also ported to the Web. Basically, we may consider that the Web content consists of several types of data such as, textual, image, audio, video, metadata, as well as, hyperlinks. Mining several types of data in an integrated system is called *Multimedia data mining* [15]. Thus, multimedia data mining is considered as a special case of WCM.

From a different point of view, Web content data maybe classified into *unstructured* data such as free text, *semi-structured* data such as HTML and XML documents, and more *structured* data such as data in tables or databases generated as HTML pages. The research done on mining free text is called Knowledge Discovery in Text (KDT) or *Text mining.*

4.2. Web Structure Mining (*WSM*)

WSM [16] tries to discover the model underlying the link structures of the Web. The model is based on the topology of the hyperlinks with or without the description of the links. This model can be used to categorize Web pages and is useful to generate information such as the similarity and relationship between different Web sites. Web structure mining is used to discover authority sites for the subjects (authorities) and overview sites for the subjects that point to many authorities (hubs). The term *Hypertext Mining* is used also in [1] for mining hypertext (text that includes hyperlinks), however, it includes mining structure as well as content.

4.3. Web Usage Mining (*WUM*)

WUM tries to make sense of the data generated by the Web surfer’s sessions or behaviors. While WCM and WSM deal with the real or primary data on the Web, WUM mines secondary data derived from the interactions of the users with the Web. Web usage data includes data from Web server logs, proxy server logs, browsers logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data that results from interactions.

4.4. Semantic Web MINING (SWM)

As the Semantic Web is based on machine processable information, therefore, automated schemes for learning the relevant information are in high demand. Semantic Web Mining aims at discovering insights about the meaning (semantics) of Web resources and their usage. Given the primarily syntactical nature of the data being mined, the discovery of *semantics* is impossible based on these data only. Therefore, formalizations of the semantics of Web sites and navigation behavior are becoming more and more common [2].

Semantic Web Mining aims at combining the two areas Semantic Web and Web Mining. This vision follows the observation that trends converge in both areas: an increasing number of researchers are working on improving the results of Web Mining by exploiting (the new) semantic structures in the Web, while others are making use of Web Mining techniques for building the Semantic Web. Last but not least, these techniques can be used for mining the Semantic Web itself. The wording Semantic Web Mining emphasizes this spectrum of possible interaction between both research areas: It can be read in two ways as: Semantic (Web Mining) and as (SemanticWeb) mining.

Following the second track, we believe that Semantic Web Mining is a fourth category that should be added to the three categories above. Also we can add *Semantics* as *metadata* that describe the data whether (Content, Structure or Usage) on the Web.

5. Web content mining

As in [12], WCM research can be categorized into two classes; from the point of view of Information Retrieval IR for unstructured and semi-structured documents, and from the point of view of Data Base (DB) for semi-structured or structured data.

5.1. Web content mining in information retrieval

Most of the research done in this area uses bags of words to represent unstructured documents (mainly free text). The bag of words or vector representation takes single words found in the training corpus as features. The representation ignores the sequence in which the words occur and is based on the statistics about single words in isolation. The features could be Boolean (a word either occurs or not in a document) or frequency based (frequency of the word in a document). Feature selection includes here removing the stop words, or using some techniques such as the information gain in order to keep the best features that identify a document or user profile. Other preprocessing steps such as stemming (getting the root of the word) may be used for the features selected.

Other feature representations may be possible such as; using information about the word position in a document, using n-grams (word sequences of length up to n) to keep the order of the words in a document, using phrases, using document concept category , using terms or using named entities such as people's names or URLs. A relational representation based on the first order logic [17] may also be used to better represent a document and formulate queries , for example, one may say that the document needed has 'word X to the left of word Y'.

In case of semi-structured documents, we notice that richer representations can be used. This is due to the additional structural (HTML and hyperlink) information in the hypertext documents. All the works in this area utilize the HTML structure inside the documents and some utilize the hyperlink structure between the documents for the document representation. The methods used are common data mining methods. The applications include hypertext classification or categorization and clustering, learning relations between Web documents, learning extraction patterns or rules and pattern discovery in semi-structured data.

5.1.1. Querying the web for resources and knowledge

A research field of IR from the World Wide Web is Web querying and the design of query languages for semi-structured data. The approach for querying structured and semi-structured documents involves the construction of tailored wrappers that map document features into instances in internal data models (i.e. graphs or tables).

WebLog [18], WebSQL [19], W3QL [20], and WebOQL [21] are all intended for information gathering from the Web. While WebLog and WebOQL aim at structuring web documents using a graph tree representations, WebSQL and W3QL are languages for finding relevant documents retrieved by several search engines in parallel.

WebML [22] is a Web mining query language, that permits resource discovery as well as knowledge discovery from a subset of the Web or the Web as a whole. WebML is an SQL-Like declarative language for Web mining. The language has primitives which allow powerful interactive querying with an OLAP (On Line Analytical Processing)–like interaction.

Example: Suppose the query is to "describe the general characteristics in relevance to authors' affiliations, publications, etc. for those documents which are popular on the Web and are on "Software Agents". A knowledge discovery query to answer this request characterized by the keyword "describe" is shown below.

```
Mine        description
In-relevance-to authors.affiliation, publication, pub_date
From        document  related-to  Computing Science
Where       one of  keywords like "software agents"
and  access-frequency = "high"
```

WebML queries are treated like information probes, being mapped to a relatively high concept layer and answered in a hierarchical manner. Moreover, the knowledge discovery power of WebML is unique as it helps to find interesting high level information about the global information base. It provides users with a high-level view of the database, statistical information relevant to the answer set, and other associative and summary information at different layers. In addition, the Multi-Layered Data Base (MLDB) model takes advantage of Web pages restructuring query languages like WebLog and available networked agents to retrieve descriptors from Web documents.

5.2. Web content mining in database systems

In the other class of Web Content Mining research, from the database point of view, we find that the database techniques on the Web are related to the problems of managing and querying the information on the Web. These problems are:

- modeling and querying the Web,
- information extraction and integration, and
- Website construction and restructuring.

Basically, the DB view tries to infer the structure of the Web site or to transform a Web site to become a database so that better information management and querying on the Web (rather than simple keywords based search) become possible. This can be achieved by finding the schema of Web documents, building a Web warehouse [23] or a Web knowledge base or a virtual database [24]. The research in this area deals with semi-structured data which is data that has some structure but no rigid schema [25].

The DB view uses representations that differ from the IR view. The DB view mainly uses the Object Exchange Model (OEM) [25] that represents semi-structured data by a labeled graph. The data in the OEM model is viewed as a graph, with objects as vertices and labels on the edges . Each object is identified by an object identifier (oid) and a value that is either atomic , such as integer, string, gif, html, etc. or complex in the form of a set of object references, denoted as a set of(label, oid) pairs.

Most of the applications in this area are; the task of *schema extraction* or discovery or *building Data Guides* [26] which is a kind of structural summary of semi-structured data. Other applications do not find the global schema but try to find frequent substructures (subschema) in semi-structured data. Another important application -which is mentioned above in IR- deals with creating a MLDB [27] in which each layer is obtained by generalizations on the lower layers and use a special purpose query language for Web mining WebML [22] to extract some knowledge from the MLDB of Web documents.

5.3. Multimedia data mining

Multimedia data mining is the mining of high-level multimedia information and knowledge from large multimedia databases. This may lead to the extraction of implicit knowledge, multimedia relationships, or other patterns not explicitly stored in multimedia databases. In fact, it may be considered as a special case of Web Content Mining.

Indeed, multimedia [28] has been the major focus for many researchers around the world, and many techniques for representing, storing, indexing, and retrieving multimedia data have been proposed. However, rare are the researchers who ventured in the multimedia data mining field. Most of the studies were confined to the data filtering step of the knowledge discovery process. The *MultiMediaMiner* [22] system is a pioneer in this area. It has been designed and developed in Simon Fraser University based on the DBMiner data mining project and the C-BIRD, a system for Content-Based Image Retrieval from Digital libraries.

The DBMiner system [29] applies multi-dimensional database structures, attribute-oriented induction, multi-level association analysis, statistical data analysis and machine learning approaches for mining different kinds of rules in relational databases and data warehouses. The C-BIRD system contains 4 major components: (i) Image Excavator (a Web agent) for the extraction of images and videos from multimedia repositories, (ii) a preprocessor for the extraction of image features and storing pre-computed data in a database, (iii) a user interface, and (iv) a search kernel for matching queries with image and video features in the database. The database used by C-BIRD contains only meta-data about the multimedia data in the image or video repository.

The MultiMediaMiner system has the following features: (i) a multidimensional multimedia data cube which is suitable for OLAP, (ii) multiple data mining modules including characterization (summarization), comparison, classification, and association, and (iii) an interactive mining interface and display.

However, the progress in research in multimedia mining is very slow due to the complexity of dealing with multimedia data specially time-based media like video and audio. Indeed, a lot of research still has to be done in this area. The work done in [30] and similar projects use mining techniques for multimedia analysis and retrieval.

6. Web structure mining

In WSM the interest is in *inter-document structure*, i.e., the structure of hyperlinks within the Web itself. This is in contrast to Web Content Mining from the DB view were the interest is the study of *intra –document structure* within Web documents.

This line of research is inspired by the study of social networks and citation analysis [1]. With social network analysis we could discover specific types of pages (such as hubs, authorities, etc.) based on the incoming and outgoing links.

Some algorithms have been proposed to model the Web topology such as HITS [16] used by the CLEVER project [31] and the Page Rank [32] used by the Google search engine. Improvements of HITS are suggested such as adding content information to the links structure [33] and outlier filtering [34]. [35] presented Parasite as another system for mining structural information on the Web.

6.1. Web structure mining applications

More *applications* of Web Structure Mining are discussed in [36] in the context of Web warehouses. These include:

- measuring the completeness of Web sites by measuring the frequency of local links that reside on the same server,
- measuring the replication of Web documents across the Web warehouse that help in identifying the mirrored sites, and
- discovering the nature of the hierarchy of hyperlinks in the Web sites of a particular domain to study how the flow of information affects the design of the Web sites.
- In general , the work done on WSM can be divided into two categories according to the approach used to mine for the knowledge:

- *using links* as discussed above in HITS and PageRank algorithms, or
- *using generalizations* to build a MLDB representation for the Web.

7. Hypertext mining

The volume of unstructured text and hypertext data exceeds that of structural data. Text and hypertext are used for digital libraries, product catalogs, reviews, newsgroups, medical reports, customer service reports, and homepages for individuals, organizations, and projects. Hypertext has been widely used long before the popularization of the Web. In this section, a summary of the work done in data mining for hypertext in general and the Web in particular is presented from a machine learning perspective[1]. Since hypertext documents consist mainly of text as well as hyperlinks, it should be clear that this work involves *hybrid* techniques that combine mining content as well as structure data from the Web, i.e. deal with both kinds of primary data that exist on the Web. It describes basic models for text, hypertext, and semi-structured data. Then it discusses techniques from supervised learning, unsupervised learning, semi-supervised learning and social network analysis.

7.1. Basic models

A model is a suitable representation for something which will suffice for our learning application.

7.1.1. Models for text

In the IR domain, documents have been traditionally represented in the *Vector Space Model* [37]. Documents are tokenized using simple syntactic rules and tokens are stemmed to canonical form (e.g., “reading” into “read “). Each canonical token represents an axis in the Euclidean space. Documents are vectors in this space. Term frequencies are also used to give weights to different terms to produce *weighted vector space* model.

Also statistical models are used such as the *Binary model.* In this model, a document is a set of terms, which is a subset of the lexicon (the universe of possible terms). Word counts are not significant but only; for each term; whether it exists in the document or not. In the *multinomial model*, one imagines a die with as many faces as there are words in the lexicon, each face t has an associated probability Pt of showing up when tossed. To compose a document, the author fixes a total word count and then tosses the die as many times, each time writing down the term corresponding to the face that shows up.

7.1.2. Models for hypertext

Hypertext has hyperlinks in addition to text. These are modeled with varying levels of detail, depending on the application. In the simplest model, hypertext can be regarded as a directed graph (D,L) where D is the set of nodes, documents, or pages, and L is the set of Links. Crude models may not need to include the text models at the nodes. More refined models will characterize some sort of joint distribution between the term distribution of a node with those in a certain neighborhood. One may also wish to recognize that the source document is in fact a sequence of terms interspersed with outbound hyperlinks. This may be used to establish specific relations between certain links and terms. For example, in [1] the authors found that Web pages about recreational bicycling lead to pages about first-aid.

7.1.3. Models for semi-structured data

Apart from hyperlinks, other structures exist on the Web, both across and within documents. One kind of inter-document structure are topic directories like the Open Directory project and Yahoo. Such services have constructed, through human effort, a giant taxonomy of topic directories. Each directory has a collection of hyperlinks to relevant and authoritative sites relevant to the specific topic. One may model Tree-structured hierarchies with anis-a (specific-topic, general-topic) relation, and an example (topic, URL) relation to assign URLs to topics. Topic directories are a special case of semi-structured data.

Semi-structured data is a point of convergence for the Web and Database communities; the former deals with *documents*, the latter with *data*. Representations for semi-structured data (such as XML) are variations on the Object Exchange Model (OEM) [38].

7.2. Hypertext mining techniques

7.2.1. Supervised learning

In supervised-learning also called Classification, the learner first receives training data in which each item is marked with a label or class from a discrete finite set. The algorithm is trained using this data, after which it is given unlabeled data and has to guess the label. Classification has numerous applications in the hypertext and semi-structured data domains. Surfers use topic directories because they help structure and restrict keyword searches. Robust hypertext classification is useful for email and newsgroup management and maintaining Web directories. An example of the work in this area is the Naïve-Bayes classifier [39].

7.2.2. Unsupervised learning

In unsupervised learning [40] of hypertext, the learner is given a set of hypertext documents, and is expected to discover a hierarchy among the documents based on the notion of similarity, and organize the documents along that hierarchy. A good clustering will collect similar documents together near the leaf levels of the hierarchy and defer merging dissimilar subsets until near the root of the hierarchy. Clustering is used to enhance search, browsing and visualization [41].

The most frequently used techniques in this category are: the basic clustering techniques (K-means clustering and Agglomerative clustering), and techniques from linear algebra such as: Latent Semantic Indexing LSI.

7.2.3. Semi-supervised learning

In real life, most often one has a relatively small collection of labeled training documents, but a larger pool of unlabeled documents. Thus learning from labeled and unlabeled documents such as in the EM algorithm [9] is referred to as semi-supervised learning.

7.2.4. Social network analysis

The Web is an example of a Social Network. Social networks are formed between papers through citation and between Web pages by hyper-linking to other Web pages. Social Network Theory is concerned with properties related to connectivity and distances in graphs, with applications like citation indexing. Starting in 1996, a series of applications of social network analysis were made to the Web graph, with the purpose of identifying the most authoritative pages related to a user query. The most important projects in this line of research are Google [32] and the HITS algorithm [16].

Google:

The Google search engine simulates a random walk on the Web graph in order to estimate *Page Rank,* which is used as a score of popularity of a Web page It is calculated as a function of the probability of jumping to a random Web page, as well as, the total number of nodes in the Web graph, and the out-degree of all the nodes (pages) in the graph that are linked to this specific Web page. Given a keyword query, matching documents are ordered by this score. Thus the popularity score is pre-computed independent of the query, hence Google can be potentially as fast as any relevance-ranking search engine.

HITS:

The Hyper-linked Induced Topic Search HITS algorithm is slightly different: it does not crawl or pre-process the Web, but depends on a search engine such as Alta Vista, which retrieves a sub-graph of the Web whose nodes (pages) match the query, as well as, the pages cited by these pages or that cite them , i.e., the inward and outward links. The algorithm then produces a set of authority pages (that contain authorized content) and a set of hub (survey) pages (that contain a list of references or links to authority pages) for the query topic. Because of the query-dependent graph construction, HITS is slower than Google. However, a variant of the HITS technique that uses a connectivity server and fetches the Web graph from it after it has pre-crawled a substantial portion of the Web is proposed in [34].

It should be noted here that social network analysis is mainly the first approach used for Web Structure mining which utilizes the linkage information in the Web and mainly represents the Web as a graph.

8. Web usage mining

Web Usage Mining (WUM) is the application of data mining techniques to discover *usage patterns* from Web data, in order to understand and better serve the needs of Web-based applications. As any knowledge discovery process, it can be divided into phases; the most important of which are; preprocessing, pattern discovery, and pattern analysis [11]. As mentioned before, the mined data in this category are the secondary data on the Web as a result of interactions. These data can range very widely but generally we could classify the data sources of usage data into three data sources: Web servers, Proxy servers, and Client browsers [11].

8.1. Usage data sources

- *Server level collection*: A Web Server Log is an important source for performing Web Usage Mining as it implicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the (possibly concurrent) access of a Web site by multiple users. These log files can be stored in various formats such as Common Log or Extended Log Formats. However, the site usage data may not be reliable due to the presence of various levels of caching within the Web environment. Cached page views are not recorded in a server log.
- *Client level collection*: Client side data collection is implemented by using a remote agent such as Java scripts or Java applets, or by modifying the source code of an existing browser to enhance its data collection capabilities. However, it requires the cooperation of the user, either enabling the functionality or to use the modified browser. It is better than server-side collection because it overcomes common problems of the former which are due to caching and session identification.
- *Proxy level collection*: A Web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy caching can be used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides. The performance of proxy caches depends on the ability to predict future page requests correctly. The usage data collected at the different sources will represent the navigation patterns of different segments of the overall Web traffic, ranging from single-user, single-site browsing behavior to multi-user, multi-site access patterns. The W3C Web Characterization Activity [42] has published a draft of Web term definitions relevant to analyzing Web usage data. The most important data abstractions defined are presented in the following subsection.

8.2. Usage data abstractions

- User: a user is defined as a single individual that is accessing a file from one or more Web servers through a browser.
- Page-View: a page view consists of every file that contributes to the display on a user's browser at one time.
- Click stream: a click-stream is a sequential series of page view requests.
- User session: is the click-stream of page views for a single user across the entire Web.
- Server session: the set of page-views in a user session for a particular Web site is referred to as the server-session or a visit.
- Episode: any semantically meaningful subset of a user or server session is referred to as an episode.

There are two approaches to WUM; the first maps usage data into relations before an adapted data mining technique is performed [43], and the other approach uses log data directly by utilizing special pre-processing techniques. Thus, the problem of data preparation (preprocessing) is very crucial to WUM.

8.3. Usage data preparation

The preprocessing of Web logs is usually complex and time demanding. It comprises four different tasks: (i) the data cleaning, (ii) the identification and the reconstruction of users_ sessions,(iii) the retrieving of information about page content and structure, and (iv) the data formatting [44].

When exploiting log information from Web servers, the major issue is the identification of users_ sessions, i.e., how to group all the users_ page requests (or click streams) so to clearly identify the paths that users followed during navigation through the web site. This task is usually quite difficult and it depends on the type of information available in log files. The most common approach is to use cookies to track down the sequence of users_ page requests. If cookies are not available, various heuristics can be employed to reliably identify users_ sessions. Note however that, even if cookies are used, it is still impossible to identify the exact navigation paths since the use of the back button is not tracked at the server level. Apart from Web logs, users_ behavior can also be tracked down on the server side by means of TCP/IP packet sniffers. Even in this case the identification of users_ sessions is still an issue, but the use of packet sniffers provides some advantages. In fact: (i) data are collected in real time; (ii) information coming from different Web servers can be easily merged together into a unique log; (iii) the use of special buttons (e.g., the stop button) can be detected in order to collect information usually unavailable in log files. Notwithstanding the many advantages, packet sniffers are rarely used in practice. Packet sniffers raise scalability issues on Web servers with high traffic, moreover they cannot access encrypted packets like those used in secure commercial transactions (through the Secure Socket Layer). Unfortunately, this limitation turns out to be quite severe when applying Web Usage Mining to e-businesses. Probably, the best approach for tracking Web usage consists of directly accessing the server application layer, but this is not always possible. First, there are issues related to the copyright of server applications. Most important, following this approach, Web Usage Mining applications must be tailored for the specific servers and have to take into account the specific tracking requirements [44].

In general, typical data mining methods [11] could be used to mine the usage data after the data have been pre-processed to the desired form. However, modifications of the typical data mining methods are also used such as in Midas [45] and [46]. Often the Web Usage Mining uses some background or domain knowledge such as navigation templates, Web content, site topology, concept hierarchies, and syntactic constraints.

8.4. Web usage mining techniques

Most of the commercial applications of Web Usage Mining exploit consolidated statistical analysis techniques. In contrast, research in this area is mainly focused on the development of knowledge discovery techniques specifically designed for the analysis of Web usage data. Most of this research effort focuses on three main paradigms: association rules, sequential patterns, and clustering [44]

Comparing association rule mining with clustering of document references it was perceived that both methods aim at identifying patterns regarding the usage of a set of web documents, while none of them offers any information about the order of documents visited. Clustering and association rule mining apparently provide the same type of results. However, there is an important difference. An association rule provides additional information about the antecedent and the consequent of a pattern, as well as about the values of confidence and support measures.

The above observations could motivate the enhancement of usage patterns produced by clustering with qualitative and quantitative information concerning antecedents and consequents, as well as about the values of confidence and support, respectively. Consequently, the combination of clustering and association rule mining techniques may potentially result into more informative and qualitative usage patterns.

8.5. Usage mining applications

The general goal of Web Usage Mining is to gather interesting information about users navigation patterns (i.e., to characterize Web users). This information can be exploited later to improve the Web site from the users_ viewpoint. The results produced by mining of Web logs can be used for various purposes: (i) to personalize the delivery of Web content; (ii) to improve user navigation through pre-fetching and caching; (iii) to improve Web design; or in e-commerce sites (iv) to improve the customer satisfaction [44].

We believe that the applications of Web Usage mining could be classified into two main approaches:

- *Applications for learning a user profile or user modeling* in adaptive (personalized) interfaces [47] , and
- *Applications for learning user navigation patterns* (impersonalized) such as in Web Miner [14], WUM [45], [48] and [49].

However, different WUM applications may be classified by their objectives into:

- *Personalization*: Personalizing the Web experience for a user is the holy grail of many Web-based applications, e.g. individualized marketing for e-commerce. Making dynamic recommendations to a Web user, based on his/her profile in addition to usage behavior is very attractive for e-commerce. The WebWatcher [50] and Letizia [51] systems are examples of this category. WebWatcher uses the technology of *software agents*. It “follows” a user as he or she browses the Web and identifies links that are potentially interesting to the user. The WebWatcher starts with a short description of a users interest. Each page request is routed through the WebWatcher proxy server in order to easily track the user session across multiple Web sites and mark any interesting links. WebWatcher learns based on the particular user’s browsing plus the browsing of other users with similar interests. This is an example of a *recommender system.* Letizia is a client side agent that searches the Web for pages similar to ones that the user has already viewed or bookmarked. The page recommendations in [52] are based on clusters of pages found from the server log for a site. The system recommends pages from clusters that most closely match the current session. Pages that have not been viewed and that are not directly linked from the current page are recommended to the user. Another personalization system not based on user intervention is presented in [53].
- *Security*: WUM can provide patterns for Web intrusion, fraud, attempted break-ins, which may be used to enhance Web site security [54]. Security is an acutely growing concern for Web-based services, especially as electronic commerce continues to grow at an exponential rate.
- *Business Intelligence*: Information on how customers are using a Web site is critical information for marketers on the Web. IBM’s SurfAid [55] provides Online Analytical Processing OLAP techniques and builds a data cube, clustering of users in addition to page view statistics.
- *System Improvement*: Performance and other service attributes are crucial to user satisfaction from services such as databases and networks. Similar qualities are expected from the users of Web services. WUM provides the key to understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission, load balancing or data distribution.
- *Site Modification*: The attractiveness of a Web site, in terms of both content and structure is crucial to many applications, e.g., a product catalog for e-commerce. WUM provides detailed feedback on user behavior, providing the Web site designer information on which to base redesign decisions. While the results of any of the projects could lead to redesigning the structure and content of a site, the adaptive Web site project [47] focuses on automatically changing the structure of a site based on usage patterns discovered from server logs. Clustering of pages is used to determine which pages should be directly linked. [56] discusses how WUM can be used for a better Web-based learning environment when the online course designer can modify his online course design after tracking and evaluating the learners’ behavior on web-based distance learning courses. In [8] a new

soft clustering technique is used for adaptive web sites.

- *Usage Characterization*: While most projects that work on characterizing the usage , content, and structure of the Web do not necessarily consider themselves to be engaged in data mining, there is a large amount of overlap between Web characterization research and Web Usage Mining [57]. In [58] a model is proposed which can be used to predict the probability distribution for various pages a user might visit on a given site.

A Meta-Web architecture is presented which is a kind of a Web knowledge base built on top of the primary data of the Web. This architecture facilitates the extraction of resources as well as knowledge from the Web using a Web (mining) query language and using an easy interface. Furthermore, the Meta-Web architecture helps to achieve the goals of "information integration" which is an interesting direction of Web Content Mining. Indeed, *information integration* was mainly concerned with integrating various databases but has changed its focus with the increasing popularity of the Web and because most of the companies are porting there data bases to the Web. This can be in the form of a Web knowledge base, or Web warehouse or in the form of a mediator.

9. Semantic web mining

The effort behind the Semantic Web is to add machine-understandable, semantic annotations to Web documents in order to access knowledge instead of unstructured material. The purpose is to allow knowledge to be managed in an automatic way. Web Mining can help to learn structures for knowledge organization (e.g., Ontologies) and to provide the population of such knowledge structures.
Research in this area can be divided into two main approaches;

- extracting semantics from the Web, and
- using semantics for Web Mining and mining the Semantic Web.

9.1. Semantic web mining applications

The first approach to semantic web mining focuses on extracting semantics from Web data. Some Applications extract the semantics created by Content such as:

- *Ontology learning:* Extracting an ontology from the Web is a challenging task. One way is to engineer the ontology by hand, but this is expensive. In [59], the expression ontology learning was coined for the semi-automatic extraction of semantics from the Web. Hence, machine learning techniques were used to improve the ontology engineering process and to reduce the effort for the knowledge engineer. Ontology learning exploits many existing resources including texts, thesauri, dictionaries, and databases, (an example is WordNet). It builds on techniques from WCM, and it combines machine learning techniques with methods from fields like information retrieval and agents, applying them to discover the 'semantics' in the data and to make them explicit. The techniques produce intermediate results which must finally be integrated in a machine-understandable format, e.g., an ontology. Mining can supplement existing (Web) taxonomies with new categories and it can also help in building new taxonomies.
- *Mapping and merging ontologies:* The growing use of ontologies leads to overlaps between knowledge in a common domain. Domain-specific ontologies are modeled by multiple authors in multiple settings. These ontologies lay the foundation for building new domain-specific ontologies in similar domains by assembling and extending multiple ontologies from repositories. The process of ontology merging takes as input two (or more) source ontologies and returns a merged ontology. Manual ontology merging using conventional editing tools without support is difficult, labor-intensive, and error-prone. Therefore, several systems and frameworks for supporting the knowledge engineer in the ontology merging task have been proposed [60]. These approaches rely on syntactic and semantic matching heuristics which are derived from the behavior of ontology engineers confronted with the task of merging ontologies. Ontology mapping is the assignment of the concepts of one ontology and their instances to the concepts of another ontology. This could be useful, for example, when one of several ontologies has been chosen as the right one for the task at hand.

The instances can simply be classified from scratch into the target ontology; alternatively, the knowledge inherent in the source ontology can be utilized by relying on the heuristic that instances from one source concept are likely to be classified together in one concept of the target ontology.

- *Instance learning:* Even if ontologies are present and users manually annotate new documents, there will still be old documents containing unstructured material. In general, the manual markup of every produced document is impossible. Also, some users may need to extract and use different or additional information from the one provided by the creator. To build the Semantic Web, it is therefore essential to produce automatic or semi-automatic methods for extracting information from Web-related documents as instances of concepts from an ontology, either for helping authors to annotate new documents or for extracting additional information from existing unstructured or partially structured documents. A number of studies investigate the use of content mining to enrich existing conceptualizations behind a Web site. For example, in [62] text categorization techniques are used to assign HTML pages to categories in the Yahoo hierarchy. This can reduce the manual effort for maintaining the Yahoo Web index.
- *Information extraction:* Information extraction from texts (IE) is one of the most promising areas of Natural Language Technologies. IE is a set of automatic methods for locating important facts in electronic documents for subsequent use. IE techniques range from the extraction of keywords from pages' text using the *term frequency inverse document frequency* (tf.idf) method known from IR, via techniques that take the syntactic structures of HTML or natural language into account, to techniques that extract with reference to an explicitly modeled target structure such as an ontology [63].

Other applications extract semantics based on structure such as:

- *Measuring page similarity and web site design:* A kind of knowledge that may be inferred from structure is the similarity between pages, useful for the popular browser application "Find similar pages". Based on the observation that pages which are frequently cited together from other pages are likely to be related, [64] proposes two algorithms for finding similar pages based on hyperlink structure. These techniques structure the set of pages, but they do not classify them into an ontology. In contrast, the hyperlink structure within pages lends itself more directly to classification. [65] proposes an ontology of page functions, where the classification of a single page with respect to this ontology can be done (semi)-automatically. For example, "navigation" pages designed for orientation contain many links and little information text, whereas "content" pages contain a small number of links and are designed to be visited for their content. This can be used to compare intended usage with actual usage. For example, a content page that is used as a frequent entry point to a site signals a challenge for site design: First, the intended entry point, which is probably the home page, should be made better-known and easier to locate. Second, additional links for navigation could be provided on the page that is currently the actual entry point. Its content may become a candidate for a new top-level content category on various "head" pages. The structure of within-page markup may also help in extracting page content: Concentrating on page segments identified by reference to the page's DOM (document object model, or tag tree) can serve to identify the main content of a page and to separate it from "noise" like navigation bars, advertisements, etc.

Other applications extract semantics based on usage such as:

- *Collaborative filtering and recommender system:* A large proportion of knowledge is socially constructed. Thus, navigation is not only driven by formalized relationships or the underlying logic of the available Web resources. Rather, it "is an information browsing strategy that takes advantage of the behavior of like-minded people". Recommender systems based on "collaborative filtering" have been the most popular application of this idea. In recent years, the idea has been extended to consider not only ratings, but also Web usage as a basis for the identification of like-mindedness ("People who liked/bought this book also looked at ...". Extracting such relations from usage can be interpreted as a kind of ontology learning, in which the binary relation "is related to" on pages (and thus concepts) is learned. A question that still has to be answered is: Can usage patterns reveal further relations to help build the Semantic Web?

9.2. Semantic web mining potentials

On the other hand, research in using semantics for Web mining and mining the semantic Web is the other face of the coin. We can point out some of the potentials of using semantics in the Web mining process.

- The types of (hyper) links are described explicitly, allowing the knowledge engineer to gain deeper insights in Web structure mining; and the contents of the pages come along with a formal semantics, allowing him to apply mining techniques which require more structured input.
- The use of ontologies as background knowledge during the preprocessing stages of the pattern discovery process such as clustering and classification leads to much better results.
- Web structure mining can be improved by taking content into account. For example, the PageRank algorithm cooperates with a keyword analysis algorithm, but the two are independent of one another. So it will consider any much cited page as ‘relevant’, regardless of whether that page’s content reflects the query. By also taking the hyperlink anchor text and its surroundings into account, the algorithm can more specifically assess the relevance for a given query.
- Ontology-based focused crawling makes use of ontologies to enhance search results as in [59].
- Web usage mining benefits from including semantics into the mining process [66] because the application expert as the end user of mining results is interested in events in the application domain, in particular user behavior, while the data available—Web server logs—are technically oriented sequences of HTTP requests. A central aim is therefore to map HTTP requests to meaningful units of application events. With the increasing standardization of many Web applications, and the increasing confluence of mining research with application domain research (e.g., marketing), the number of standard courses of events is likely to grow which makes this area of very high importance.

10. Web mining and the agent paradigm

Web mining is often viewed from or implemented within a software agent paradigm. Thus Web mining has a close relationship with the software intelligent agents. Indeed, some software agents for example, [67] and [66] perform data mining tasks in order to achieve their goals.

Although Software Intelligent Agents (IA) is a rapidly developing area of research, there is no universal definition of IA. One of the definitions that may illustrate the functionality of an IA is:

“*Intelligent agents* continuously perform three functions: perception of dynamic conditions in the environment; action to affect conditions in the environment; and reasoning to interpret perceptions, solve problems, draw inferences and determine actions” [68].

Intelligent agents have several features some of which are necessary [69] such as: Autonomy; agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state , and Adaptability; an agent is able to improve over time, i.e., becomes better at achieving its goals with experience. An optional feature of Intelligent Agents which is needed for the multi-agents paradigm is: Collaboration; which is the ability to collaborate with other agents to achieve their goals, so that they have the ability to interact with other agents and possibly humans via some agent communication language. This feature is sometimes thought to be mandatory based on the nature of the application.

Different types of agents exist and different taxonomies for them have been suggested according to different criteria. If the application is the base of our taxonomy then we can find that some subcategories of intelligent agents have a close relationship with Web mining tasks. The sub-categories of software agents that are relevant for web mining tasks are User Interface Agents, and Distributed agents.

User interface agents [70] try to maximize the productivity of current users' interaction with the system by adapting behavior. Those that can be classified into the Web mining agent category are: Information Filtering Agents, Information Retrieval Agents, and Personal Assistance Agents.

For example, ARCH [71] is an Adaptive agent for Retrieval based on Concept Hierarchies which is considered an information retrieval agent. It is a client side agent, for assisting users in one of the most difficult IR tasks which is formulating an effective search query based on a modular concept hierarchy and a learned user profile.

Distributed agents mainly Multi-agent systems use web mining tasks to perform their jobs which usually needs collaboration such as Collaborative Filtering. [72] presents techniques for improving multi-agents for mining and searching the Web. [67] presents a Web mining system based on multi-agents.

There are two frequently used methods for developing intelligent agents based on machine learning techniques which are: the content-based approach and the collaborative approach.

10.1. The content-based approach

In this approach, for instance, in order to use agents in text filtering the system searches for items similar to those the user prefers based on a comparison of content.

For example, the WebWatcher [50] system helps users locate information on the Web by taking keywords from users, suggesting hyperlinks, and receiving evaluation. Then it also lets users get additional similar documents. Furthermore, the Personal WebWatcher system, is a content-based intelligent agent that uses text-learning for user customized Web browsing. It is a personal assistant agent for Web browsing that accompanies the user from page to page and highlights interesting hyperlinks. It generates a user profile based on the content analysis of the requested pages without requesting any keywords or ratings from the user.

The work done in the field of text-learning intelligent agents may be compared based upon three key criteria which are: what representation the particular application uses for documents, how it selects features, and what learning algorithm it uses. These systems assist users by finding information or performing some simple tasks on their behalf [70]. This system might help in Web browsing by retrieving documents similar to already requested ones.

The Lira system [73] learns to browse the Internet on a user's behalf. It searches the Web by taking a bounded amount of time, selecting the best pages and receiving an evaluation from the user. Lira uses the evaluation to update the search and selection heuristics.

The Letizia project at MIT, is also a user interface agent for assisting Web browsing. It does not require any keywords or rating from the user because it infers the user's interests from his browsing behavior. While the user is reading a document, Letizia performs a breadth first search from the current document then it suggests potentially interesting hyperlinks found during the search to the user in a separate browser window.

The content based approach has its roots in information retrieval but it has difficulties in capturing different aspects of content such as music, movies and images. Consequently the collaborative approach was suggested to help in recommender systems.

10.2. The collaborative approach

In contrast to the content-based approach which can be successfully applied to a single user, the collaborative approach assumes that there is a set of users using the system. In the collaborative approach (sometimes called social learning), advice to the user is based on the reaction of other users. The system searches for users with similar interests and recommends the items these users liked. Instead of computing the similarity between items (documents, music,…etc..), the system calculates the similarity between users as discussed in Sec 9.2.

The collaborative approach is usually suitable for non-text data (movies, music..etc..) but there are also systems that use it on text data. These systems are called recommender systems [74].

Siteseer [75], is a Web-page recommendation system, that uses an individual's bookmarks and the organization of bookmarks within folders for predicting and recommending relevant pages. The system measures the degree of overlap (such as common URLs) between the bookmark files of different users and then groups users according to that similarity. When making recommendations, Siteseer gives priority to URLs obtained from similar folders and URLs that appear in bookmark files of similar users.

My spider [76] is a threaded multi-agent system designed for information discovery. It complements the search engine with web mining capabilities.

The Collaborative Spider [77] is a multi agent system designed to provide post retrieval analysis and enable across user collaboration in web search and mining. The system allows the user to annotate search sessions and share them with others.

The FAB system [74] for Web document recommendation combines content-based and collaborative approaches. It uses the content-based approach to generate a profile that represents a single user's interests and uses the collaborative approach to find similar users. The user's ratings are used to update that person's personal profile. The two approaches are combined, and the pages matching the user's profile as well as the pages highly rated by similar users are recommended. FAB measures the similarity between users by the similarity of their profiles.

11. Summary and research directions

Indeed, the World Wide Web (Web) presents new challenges to the traditional data mining algorithms that work on flat data and were mainly designed for structured data in databases. Hence, some of the traditional data mining algorithms have been extended and new algorithms have been designed to work on the Web data. Moreover, the evolution of the Semantic Web as the second generation WWW has reinforced the research in Web mining and added much more potentials to it.

This paper tries to present and classify the tremendous effort done in the area of Web mining and establishing the Semantic Web. This field of research integrates research in different fields: Information Retrieval, Database, Machine Learning, Pattern Recognition and Multi-Agent Systems. We believe that Semantic Web Mining should be added to the taxonomy of Web Mining Techniques as a fourth category besides Web Content Mining, Web Structure Mining and Web Usage Mining. Thus, the Web mining taxonomy is revisited and extended. Then, each Web mining category is discussed pointing out its design issues, representation problems, learning algorithms used and major applications. Furthermore, the close relationship between Web Mining and the Software Agent paradigm is illustrated. In fact, different categories of Software Intelligent Agents use Web mining techniques to achieve their goals on the Web and inversely Software agents are used to aid in Web Mining techniques.

Research directions in this area are very wide as is the field. We can point out some of them as can be inferred from the previous discussions.

The area of *multimedia data mining* can benefit a lot of combined techniques for Web content Mining and Semantic web mining. Indeed further research is needed for extracting semantics out of multimedia data such as images and video in order to build the semantic web by annotating multimedia data and building ontologies for different multimedia data types.

New representation models, and learning algorithms can be developed for *XML* documents which are now replacing HTML with added structure and better semantics. Thus, the area of *XML Mining* is a fertile area of research.

An interesting fact is that graph structures occur almost everywhere in Web mining research. There are many opportunities for (existing or new) machine learning algorithms that could work with these representations or that could take advantage of the available structures on the Web taking into consideration the huge volume of data on the Web. Thus the scalability of well known graph algorithms such as graph partitioning or clustering techniques has to be revised and adapted for the large number of nodes available on the Web.

Future research on semantic web mining is expected to be done in an iterative fashion, i.e., extracting semantics from the Web, integrating the mined knowledge with the mined data on the Web and then utilizing it again for better mining in the future. Thus the Web will be continuously evolving towards a "better Web" which will be reflected from its usage. A lot of experimental work is needed to evaluate this research and new Web metrics and measurements are needed. Research in Human Computer Interaction can help in this area to evaluate Web sites usability at different stages of its evolution.

The exploitation of the Semantic web potentials discussed in the paper may lead to much better mining of the Web, such as for example, using ontologies for enhancing clustering techniques such as in conceptual clustering. The machine learning algorithms which are used in the pattern discovery process can be revisited exploiting the knowledge embedded in the semantic web.

Exploiting web mining techniques for detecting security threats is a very important area of research. Some research issues such as recognizing frauds, characterizing them and detecting new ones are highly in demand.

Also, protecting privacy of organizations and users during the mining process is an important issue. There is a need to develop approaches, methodologies and tools that can be used to verify and validate that a Web service is using an organization information in a manner consistent with its stated policy.

References

[1] S. Chakrabarti, "Data Mining for Hypertext: A Tutorial Survey", SIGKDD Explorations, Vol. 1 (2), pp. 1-11, Jan. (2000).

[2] Gerd Stumme, Andreas Hotho and Bettina Berendt, "Semantic Web Mining: State of the Art and Future Directions", Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 4, pp. 124-143 (2006).

[3] P. Patel-Schneider and D. Fensel, "Layering the Semantic Web: Problems and Directions", In: Proceedings First International Semantic Web Conference, Vol. 2342. of LNCS, Springer, pp. 16-29 (2002)

[4] T.R. Gruber, Towards Principles for the Design of Ontologies Used for Knowledge Sharing, in: N. Guarino, R. Poli (Eds.), Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer, Deventer, Netherlands (1993).

[5] O.Etzioni, "The World Wide Web: Quagimare or Gold to Mine, Communications of the ACM Journal, Vol. 39 (11), pp. 65-68 (1996).

[6] U. Fayyad, G. Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview", in: Proc. ACM SIGKDD (1994).

[7] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", in: Proc. of the 20th VLDB Conf., 487-499, Santiago, Chile, (1994).

[8] R.A. Shokry, Amani A. Saad, N.M. El-Makkey and Mohamed A. Ismail, "Using New Soft Clustering Technique in Adaptive Web Site", IAT Workshops, pp. 281-286 (2006).

[9] T. Mitchell, (ed.), Machine Learning, McGraw Hill (1997).

[10] Xu Yusheng; Ma Zhixin, Li Lian and T.S. Dillon, "Effective Pruning Strategies for Sequential Pattern Mining, Proceedings of the First Int'l Workshop on Knowledge Discovery and Data Mining", WKDD 2008. Vol. 1, Issue, 23-24, pp. 21–24, Jan. (2008).

[11] Srivastava, Cooley, Deshpande, and Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1 (2), pp. 12-23, Jan. (2000).

[12] R. Kosala and H. Blockeel, "Web Mining Research: A Survey", ACM SIGKDD Explorations, Vol. 2 (1), pp. 1-15, July (2000).

[13] S. Madria, S. Bhowmick, W. Ng and E-P. Lim, "Research Issues in Web Data Mining", in: Proc. of Data Warehousing and Knowledge Discovery, 1st Int'l Conf. DaWaK'99, 303-312 (1999).

[14] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", in: Proceedings of the 9th Int'l Conf. On Tools with Artificial Intelligence (ICTAI'97), Nov (1997).

[15] O. Zaiane, J. Han, Z. Li, S. Chee, and J. Chiang, "MultiMedia Miner: A System Prototype for MultiMedia Data Mining", ACM SIGMOD Int'l Conf. 98, Seattle, June (1998).

[16] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", in: Proc. 9th ACM-SIAM Symposium on Discrete Algorithms (1998).

[17] M. Craven et al., "Learning to extract Knowledge from the World Wide Web", in: Proc. of the 5th National Conf. on Artificial Conf. on Artificial Intelligence (AAAI98), pp. 509-516 (1998).

[18] L. Lakshmanan, F. Sadri and I. Subramanian, "A Declarative Language for Querying and Restructuring the Web", in: Proceedings of the 6th Intl. Workshop on Research Issues in Data Engineering (RIDE'96), pp. 12-21 (1996).

[19] Mendelzon, G. Mihaila and T. Milo, "Querying the World Wide Web", in: Proc. of the 4th Int'l Conf. on Parrallel and Distributed Information Systems, pp. 80-91 (1996).

[20] D. Konopnicki and O. Shmueli, "W3QS: A Query System for the World Wide Web", In: Proc. 21st VLDB Conf., Zurich, Switzerland (1995).

[21] G. Arocena and A. Mendelzon, "WebOQL: Restructuring Documents, Databases and Webs", in: Proc. of ICDE Conf., Orlando, Florida, USA, February (1998).

[22] O. Zaiane and J. Han, "Web ML: Querying the World Wide Web for Resource and Knowledge", in: Proc. of ACM CIKM'98 Workshop on Web Information and Data Management (WIDM'98), pp. 9-12, Washington DC (1998).

[23] O. Zaiane, M. Xin and J. Han, "Discovering Web Access Patterns and Trends by applying OLAP and Data Mining Technology on Web Logs", in: Proc. Advances in Digital Libraries (ADL'98), Santa Barbara, Apr. (1998).

[24] O. Zaiane, "Building Virtual Web Views", in: The Special Issue on Warehouse Design for Structured and Semi-structured Data, of The Journal of Data and Knowledge, Vol. 39 (2), pp. 143-163, Nov. (2001).

[25] S. Abiteboul, D. Quass, J. McHugh, J. Widom and J. Wiener, "The Lorel Query Language for Semi-Structured Data", Int'l Journal on Digital Libraries, Vol. 1 (1), pp. 68-88 (1997).

[26] R. Goldman and J. Widom, "Dataguides: Enabling Query Formulation and Optimization in Semi-Structured Databases", in: Proc. of the 23rd Int'l Conf. on Very Large Data BasesVLDB'97, Athens, Greece, pp. 436-445 (1997).

[27] J. Han, O. Zaiane and Y. Fu, "Resource and Knowledge Discovery in Global Information Systems: A Multiple Layered Database Approach", in: Proc. of Conf. on Advances in Digital Libraries, Washington DC, May (1995).

[28] S. Khoshafian and A. Baker, (eds.), Multimedia and Imaging Databases, Morgan Kaufmann Publishers (1996).

[29] J. Han et al. "DBMiner: A System for Data Mining in Relational Databases and Data Warehouses", in: Proc. CASCON'97: Meeting of Minds, 249-260, Toronto, Canada, Nov. (1997).

[30] Apostol Natsev, John R. Smith, Jelena Te?i?, Lexing Xie and Rong Yan, "IBM Multimedia Analysis and Retrieval System," ACM International Conference on Image and Video Retrieval (ACM CIVR), Niagara Falls, Ontario, Canada, July (2008).

[31] Cakrabarti et al., "Hyper Searching the Web", Scientific American Journal, June (1999).

[32] S. Brin and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", in: Proc. of the 7th Int'l WWW Conf., Brisbane, Australia (1998).

[33] S. Chakrabarti et al., "Automatic Resource List Compilation by Analyzing Hyperlink Structure and Associated Text", in: Proc. 7th Int'l World Wide Web Conf. (1998).

[34] K. Bharat and M. Henzinger, 'Improved Algorithms for Topic Distillation in a

Hyperlinked Environment", in: Proc. of the 21st annual int'l ACM SIGIR Conf. on Research and Development in Information Retrieval , August 24-28, pp. 104-111, Melbourne, Australia (1998).

[35] E. Spertus, "Parasite: Mining Structural Information on the Web. Computer Networks and ISDN Systems", The Int'l Journal of Computer and Telecommunication Networking, Vol. 29, pp. 1205-1215 (1997).

[36] S. Madria, S. Bhowmick, W. Ng and E-P. Lim, "Research Issues in Web Data Mining", in: Proc. of Data Warehousing and Knowledge Discovery, 1st Int'l Conf. DaWaK'99, pp. 303-312 (1999).

[37] G. Salton and M. McGill, (eds), Introduction to Modern Information Retrieval, McGraw Hill (1983).

[38] J. McHugh, S. Abiteboul, R. Goldman, D. Quass and J. Widom, "Lore: A Database Management System for Semistructured Data", SIGMOD Record, Vol. 26 (3), pp. 54-66, Sept. (1997).

[39] S. Chakrabarti, B. Dom, R. Agrawal and P. Raghavan, "Using Taxonomy, Discriminants, and Signatures to navigate in Text Databases", in: VLDB, Athens, Greece, Sept. (1997).

[40] Jain and R. Dubes, Algorithms for Clustering Data, Prentice Hall, 1988.

[41] M. Hearst and C. Karadi, "Cat-a Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy", in: Proc. of the 20th Annual Int'l ACM/ SIGIR Conf., Philadelphia, PA, July (1997).

[42] World Wide Web Committee Web Usage Characterization Activity. http://www.w3.org/WCA

[43] M. Koutri, N. Avouris and S. Daskalaki, "A Survey on Web Usage Mining Techniques for Web-Based Adaptive Hypermedia Systems", in: S. Y. Chen and G.D. Magoulas (ed), Adaptable and Adaptive Hypermedia Systems, Chapter 7, IRM Press, pp. 125-149, Hershey (2005).

[44] Federico Michele Facca, Pier Luca Lanzi, "Mining Interesting Knowledge from Web-Logs: A Survey", Data and Knowledge Engineering, Vol. 53 (3), pp. 225 – 241, June (2005).

[45] M. Spiliopoulou, "Web Usage Mining for Web Site Evaluation", Communications of the ACM Journal, Vol. 43 (8), pp.127-134, August (2000).

[46] B. Masand and M. Spiliopoulou, "Report on WebKDD-99: Workshop on Web Usage Analysis and User Profiling", ACM SIGKDD Explorations, Vol. 1 (2) (2000).

[47] M. Perkowitz and O. Etzioni, "Adaptive Web Sites: Automatically Synthesizing Web Pages", in: 15th National Conference on Artificial Intelligence, Madison, WI (1998).

[48] M. Buchner, M. Baungarten, M. Anand, S. Mulvena and J. Hughes, "Navigation Pattern Discovery from the Internet Data", in: Proc. of WebKDD'99, San Diego, CA, August (1999).

[49] Bernandete Riberio and Alberto Cardoso, "Behavior Pattern Miming During the Evaluation Phase in an E-Learning Course", International Conference on Engineering Education, Purtugal, Sept. (2007).

[50] T. Joachims, D. Freitag and T. Mitchell, "WebWatcher: A Tour Guide for the World Wide Web", in: Proc. of the Int'l Joint Conf. on Artificial Intelligence IJCAI-97, pp. 770-777 (1997).

[51] H. Leiberman, "Letizia: An Agent that Assists Web Browsing", in: Proc. of the Int'l Joint Conf. on AI, Montreal, Canada (1995).

[52] B. Mobasher, R. Cooley and J. Srivastava, "Creating Adaptive Web Sites through Usage-Based Clustering of URLs", in: Knowledge and Data Engineering Workshop (1999).

[53] R. Baraglia, F. Silvestri, "Dynamic Personalization of Web Sites without User Intervention", In Communication of the ACM, Vol. 50 (2), pp. 63-67 (2007).

[54] IJ Jian, Guo-Yin, Guo-Chang and Li. Jian , "The Design and Implementation of Web Mining in Web Site Security", Journal of Marine Science and Application, Vol. 1 (2), June (2003).

[55] Surfaid Analytics, http://Surfaid.dfw.ibm.com

[56] O. Zaiane, "Web Usage Mining for a Better Web-Based Learning Environment", in: Proc. of Conf. On Advanced Technology for Education", Banff, Alberta, pp. 60-64, Jun 27-28 (2001).

[57] J. Borges and M. Levene, "A Fine Grained Heuristics to Capture Web Navigation Patterns", ACM SIGKDD Explorations, Vol. 2 (1), pp. 40-50, July (2000).

[58] B. Huberman, P. Pirolli, J. Pitkow and R. Kukose, "Strong Regularities in World Wide Web Surfing", Technical Report, Xerox PARC (1998).

[59] Maedche and S. Staab, "Ontology Learning for the Semantic Web", IEEE Intelligent Syst. Vol.16 (2), pp.72-79 (2001).

[60] D. McGuinness, R. Fikes, J. Rice and S. Wilder, An Environment for Merging and Testing Large Ontologies, in: Proceedings of the Seventh Inter-national Conference on Principles of Knowledge Representation and Reasoning (KR2000), Breckenridge, Colorado, USA, pp. 483-493 (2000).

[61] D. Zhang and W.S. Lee, "Learning to Integrate Web Taxonomies", Journal of Web Semantics, Vol. 2 (2), pp. 131-151 (2004).

[62] D. Mladenic, "Turning Yahoo to Automatic Web-page Classifier", In European Conference on Artificial Intelligence, pp. 473-474 (1998).

[63] A.H.F. Laender, B.A. Ribeiro-Neto, A.S. da Silva and J.S. Teixeira, "A Brief Survey of Web data Extraction Tools", SIGMOD Record, Vol. 31 (2), pp. 84- 93 (2002).

[64] J. Dean and M.R. Henzinger, "Finding Related Pages in the World Wide Web", in: Proceedings of the Eighth International World Wide Web Conference WWW-1999, Toronto (1999).

[65] R. Cooley, B. Mobasher and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", Knowledge and Information systems Journal, Vol. 1 (1), pp. 1-26 (1999).

[66] Kiavash Bahreini and Atilla Elci, "SDISSASA: A Multiagent-Based Web Mining via Semantic Access to Web Resources in Enterprise Architecture", Computer Software and Applications Conference, Annual International, 2008 32nd Annual IEEE International Computer Software and Applications Conference, pp. 553-558 (2008).

[67] W. Hu and B. Meng, "Design and Implementation of Web Mining System Based on Multi-Agent", in: Li, Wang and Dong (eds), ADMA 2005, LNAI 3584, pp. 491-498, Springer-Verlag , Berlin Heidelberg (2005).

[68] Barbara Hayes-Roth, "An Architecture for Adaptive Intelligent Systems", Artificial Intelligence, Vol. 72 (1-2), pp. 329-365 (1995).

[69] H. Nwana, "Software Agents: An Overview", Knowledge Engineering Review, Vol. 11 (3), pp. 205-244, October/November (1996).

[70] D. Mladenic and J. Stefan, "Text Learning and Related Intelligent Agents: A Survey", IEEE Intelligent Systems Journal, Vol. 14 (4), pp. 44-54 (1999).

[71] S. Parent, B. Mobasher and S. Lytinen, "An Adaptive Agent for Web Exploration Based on Concept Hiearchies", in: Proc. of the 9th International Conference. On Human Computer Interaction, New Orleans, August (2001).

[72] E. Videma, C. Porcel, F. Herrera, L. Martinez and A. Lopez-Herrera, "In: Techniques to Improve Multi-Agent Systems for Searching and mining the Web", Springer-Verlag, pp. 463-486 (2005).

[73] M. Balabanovi'c and Y. Shoham, "Learning Information Retrieval Agents: Experiments with Automated Web Browsing", AAAI 1995 Spring Symposium on Information Gathering from Heterogeneous Distributed Environments, AAAI Press, Menlo Park, Calif (1995).

[74] M. Balabanovi'c and Y. Shoham, "Fab: Content - Based Collaborative Recommendation", Communications of the ACM, Vol.40 (3), pp. 66-70 (1997).

[75] J. Rucker and J. Marcos, "Siteseer, Personalized Navigation for the Web",

Communications of the ACM, Vol. 40 (3), pp. 73-75 (1997).

[76] Filippo Menczer, "Complementing Search Engines with Online Web Mining Agents, Decision Support Systems Journal, El Sevier Publishing Company, 992 (2002).

[77] M. Chau, D. Zang, H. Chen, M. Huang, and D. Hendriawan, "Design and Evaluation of a Multi-Agent Collaborative Web Mining System", Decision Support System Journal, Elsevier Science, Vol. 35, pp. 167-183 (2003).

Received May 3, 2009
Accepted June 30, 2009