Web Mining Research: A Survey

Raymond Kosala
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
Raymond@cs.kuleuven.ac.be

Hendrik Blockeel
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
Hendrik.Blockeel@cs.kuleuven.ac.be

ABSTRACT

With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. The Web mining research is at the cross road of research from several research communities, such as database, information retrieval, and within AI, especially the sub-areas of machine learning and natural language processing. However, there is a lot of confusions when comparing research efforts from different point of views. In this paper, we survey the research in the area of Web mining, point out some confusions regarded the usage of the term Web mining and suggest three Web mining categories. Then we situate some of the research with respect to these three categories. We also explore the connection between the Web mining categories and the related agent paradigm. For the survey, we focus on representation issues, on the process, on the learning algorithm, and on the application of the recent works as the criteria. We conclude the paper with some research issues.

Keywords

Web, data mining, information retrieval, information extraction

1. INTRODUCTION

The World Wide Web (Web) is a popular and interactive medium to disseminate information today. The Web is huge, diverse, and dynamic and thus raises the scalability, multimedia data, and temporal issues respectively. Due to those situations, we are currently drowning in information and facing information overload [86]. Information users could encounter, among others, the following problems when interacting with the Web:

a. Finding relevant information: People either browse or use the search service when they want to find specific information on the Web. When a user uses search service he or she usually inputs a simple keyword query and the query response is the list of pages ranked based on their similarity to the query. However today's search tools have the following problems [23]. The first problem is low precision, which is due to the irrelevance of many of the search results. This results in a difficulty finding the relevant information. The second problem is low recall, which is due to the inability to index all the information available on the Web. This results in a difficulty finding the unindexed information that is relevant. See [81] for some other search engine problems.

b. Creating new knowledge out of the information available on the Web: Actually this problem could be regarded as a sub-problem of the problem above. While the problem above is usually a query-triggered process (retrieval oriented), this problem is a data-triggered process that presumes that we already have a collection of Web data and we want to extract potentially useful knowledge out of it (data mining oriented). Recent research [34; 85; 29] focuses on utilizing the Web as a knowledge base for decision making.

c. Personalization of the information: This problem is often associated with the type and presentation of information, since it is likely that people differ in the contents and presentations they prefer while interacting with the Web.

On the other hand, the information providers could encounter these problems, among others, when trying to achieve their goals on the Web:

d. Learning about consumers or individual users: This is a problem that specifically deals with the problem c above, which is about knowing what the customers do and want. Inside this problem, there are sub-problems such as mass customizing the information to the intended consumers or even to personalize it to individual user, problems related to effective Web site design and management, problems related to marketing, etc.

Web mining techniques could be used to solve the information overload problems above directly or indirectly. However, we do not claim that Web mining techniques are the only tools to solve those problems. Other techniques and works from different research areas, such as database (DB), information retrieval (IR), natural language processing (NLP), and the Web document community, could also be used. By the direct approach we mean that the application of the Web mining techniques directly addresses the above problems. For example, a Newsgroup agent that classifies whether the news is relevant to the user. By the indirect approach we mean that the Web mining techniques are used as a part of a bigger application that addresses the above problems. For example, Web mining techniques could be used to create index terms for the Web search services.

The Web mining research is a converging research area from several research communities, such as database, IR, and AI research communities especially from machine learning and NLP. This paper is an attempt to put the research done in a more structured way from the machine learning point of view. However, the methods of the research that we survey do not necessarily use well-known machine learning algorithms. Since this is a huge, interdisciplinary, and very dynamic research area, there are undoubtedly some omissions in our coverage.

This paper is structured as follows. In section 2 we give an overview of Web mining, describe some confusions in the usage of the term Web mining, provide a classification, and relate this classification to the agent paradigm. In section 3, 4 and 5 we describe some research that represent the range of the research in their respective categories. In section 6 we discuss some related work and finally we conclude in section 7.

2. WEB MINING

2.1 Overview

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services [41]. This area of research is so huge today partly due to the interests of various research communities, the tremendous growth of information sources available on the Web and the recent interest in e-commerce. This phenomenon partly creates confusion when we ask what constitutes Web mining and when comparing research in this area. Similar to Etzioni [41], we suggest decomposing Web mining into these subtasks, namely:

1. Resource finding: the task of retrieving intended Web documents.
2. Information selection and pre-processing: automatically selecting and pre-processing specific information from retrieved Web resources.
3. Generalization: automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. Analysis: validation and/or interpretation of the mined patterns.

By resource finding we mean the process of retrieving the data that is either online or offline from the text sources available on the Web such as electronic newsletters, electronic newswire, newsgroups, the text contents of HTML documents obtained by removing HTML tags, and also the manual selection of Web resources. We also include text sources that originally were not accessible from the Web but are accessible now, such as online texts made for research purposes only, text databases, etc. The information selection and pre-processing step is any kind of transformation processes of the original data retrieved in the IR process. These transformations could be either a kind of pre-processing that are mentioned above such as removing stop words, stemming, etc. or a pre-processing aimed at obtaining the desired representation such as finding phrases in the training corpus, transforming the representation to relational or first order logic form, etc. In step 3 above, machine learning or data mining techniques are typically used for the generalization. We should also note that humans play an important role in the information or knowledge discovery process on the Web since the Web is an interactive medium. This is especially important for validation and/or interpretation in step 4. Thus, interactive query-triggered knowledge discovery is as important as the more automatic data-triggered knowledge discovery. However, we exclude the knowledge discovery done manually by humans. As we will see later in section 3, the process 1 - 3 - 4 is also used.

Thus, Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the Web data. It implicitly covers the standard process of knowledge discovery in databases (KDD) [43]. We could simply view Web mining as an extension of KDD that is applied on the Web data. From the KDD point of view, the information and knowledge terms are interchangeable [44]. There is a close relationship between data mining, machine learning and advanced data analysis [90]. However, throughout the paper, we discuss the Web mining research where machine learning techniques are used. Although mining is an intriguing word to use, it is not a good metaphor to describe the overall knowledge discovery process [44] and what people really do in the field [62]. Web mining is often associated with IR or IE. However, web mining or information discovery on the Web is not the same as IR or IE.

2.1.1 *Web Mining and Information Retrieval*

Some have claimed that resource or document discovery (IR) on the Web is an instance of Web (content) mining and others associate Web mining with intelligent IR. Actually IR is the automatic retrieval of all relevant documents while at the same time retrieving as few of the non-relevant as possible [119]. IR has the primary goals of indexing text and searching for useful documents in a collection and nowadays research in IR includes modeling, document classification and categorization, user interfaces, data visualization, filtering, etc. [10]. The task that can be considered to be an instance of Web mining is Web document classification or categorization, which could be used for indexing. Viewed in this respect, Web mining is part of the (Web) IR process. However, we should note that not all of the indexing tasks use data mining techniques.

2.1.2 *Web Mining and Information Extraction*

IE has the goal of transforming a collection of documents, usually with the help of an IR system, into information that is more readily digested and analyzed [33]. IE aims to extract relevant facts from the documents while IR aims to select relevant documents [100]. While IE is interested in the structure or representation of a document, IR views the text in a document just as a bag of unordered words [123]. Thus, in general IE works at a finer granularity level than IR does on the documents. However, the differences between the two become blurred if the interest of IR is in extraction [101], and when used in the context of vague forms of information in which a full text IR system can provide some IE features [123].

Building IE systems manually is not feasible and scalable for such a dynamic and diverse medium such as Web contents [94]. Due to this nature of the Web, most IE systems focus on specific Web sites to extract. Others use machine learning or data mining techniques to learn the extraction patterns or rules for Web documents semi-automatically or automatically [77]. Within this view, Web mining is part of the (Web) IE process. Other views regarding the relationship between (Web) IE and Web mining also exist. The results of the IE process could be in the form of a structured database or could be a compression or summary of the original text or documents. One could view for the former that IE is a kind of pre-processing stage in the Web mining process, which is the step after the IR process and before the data mining techniques are being performed. In a similar view, IE can also be used to improve the indexing process, which is part of the IR process. Conversely, one could also argue for the latter that IE is an instance of text or Web mining since the summary or the compressed form of a document is a new form of information that does not exist before. However, we advocate the view that Web mining is used to improve Web IE (Web mining is part of IE).

There are basically two types of IE: IE from unstructured texts and IE from semi-structured data [93]. There are considerable differences between the IE systems that are used for unstructured documents with those that are used for semi-structured or even structured documents. IE tasks from unstructured natural language texts (classical or traditional IE tasks) typically use a rather basic to a slightly deeper linguistic pre-processing before performing data mining. Classical or traditional IE research, with roots on the NLP community, has been studied for quite a long time [33; 123]. We could say that Advanced Research Projects Agency (ARPA) helped creating the field (classical IE) because the evaluations of IE cannot be separated from the ARPA sponsored Message Understanding Conferences (MUCs) and the TIPSTER IE project [123; 7]. MUCs and TIPSTER are competitive environments that seek to improve IE and IR technologies [33; 22]. Classical IE usually relies on linguistic pre-processing such as syntactic analysis, semantic analysis, and discourse analysis [111; 93; 77]. Indeed, classical IE could be called a core language technology [123].

With the increasing popularity of the Web, there is a need for structural IE systems that extract information from semi-structured documents. Structural IE research is different from the classical one as it usually utilizes the meta-information (e.g. HTML tags [111], simple syntactics [77], or delimiters [93] that are available inside the semi-structured data. Structural IE approaches that do not use linguistic constraints are termed wrapper induction [93]. Some of the structural IE systems are built manually by knowledge engineering approach, for examples see [26; 9; 60]. However, more and more structural IE systems for the Web are built (semi-) automatically using machine learning techniques or other algorithms as building the systems manually is no longer appropriate [77]. Some examples are [78; 52; 67; 94; 59; 111]. These systems are usually built by using machine learning or data mining techniques, which learn extraction rules from the annotated corpora. For more explanations and the categories of IE we point interested readers to the following survey papers. For classical IE and the issues of IE for unstructured texts we refer to [33; 22; 7; 111] and for structural IE we refer to [111; 93].

2.1.3 *Web Mining and Machine Learning Applied on the Web*

Web mining is not the same as learning from the Web or machine learning techniques applied on the Web. On the one hand, there are some applications of machine learning applied on the Web that are not instances of Web mining. An example of this is a machine learning technique that is used to spider the Web efficiently for a specific topic [105; 88] that emphasize on planning the best path that is going to be traversed next. On the other hand, there are some other methods used for Web mining besides machine learning methods. Some examples are some proprietary algorithms that are used for mining the hubs and authorities [24], DataGuides [56; 57] and Web schema discovery [120; 97]. However, there is a close relationship between the two research areas. Machine learning techniques support and help Web mining as they could be applied to the processes in Web mining. For example recent research [91] shows that applying machine learning techniques could improve the text classification process compared to the traditional IR techniques. In short, Web mining intersects with the application of machine learning on the Web.

2.2 Web Mining Categories

In this section we give the overview of each category. More detailed explanations are given in the respective sections. Similar to Madria, et al. [85] and Borges and Levene [16], we categorize Web mining into three areas of interest based on which part of the Web to mine: Web content mining, Web structure mining, and Web usage mining. Web content mining describes the discovery of useful information from the Web contents/data/documents. However, what consist of the Web contents could encompass a very broad range of data. Previously the Internet consists of different types of services and data sources such as Gopher, FTP and Usenet. Now most of those data are either ported to or accessible from the Web. It is mentioned in [66] that in the last several years the growth in the amount of government information has been tremendous. We also know the existence of Digital Libraries that are also accessible from the Web. We also see that many companies are transforming their businesses and services electronically. As a consequence many of the company databases that previously resided in the legacy systems are being ported to or made accessible from the Web. Thus the employees, partners, or even customers could access some of the company database directly from Web based interfaces. Another consequence of this transformation is the existence of Web applications so that the users could access the applications through Web interfaces. Many applications and systems are being migrated to the Web and many types of applications are emerging in the Web environment. Of course some of the Web content data are hidden data, which cannot be indexed. These data are either generated dynamically as a result of queries and reside in the DBMSs or are private. In short, the Web already contains many kinds and types of data.

Basically, the Web content consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks. Recent research on mining multi types of data is termed multimedia data mining [128]. Thus we could consider multimedia data mining as an instance of Web content mining. However this line of research still receives less attention than the research on the text or hypertext contents [128; 90]. The Web content data consist of unstructured data such as free texts, semi-structured data such as HTML documents, and a more structured data such as data in the tables or database generated HTML pages. However, much of the Web content data is unstructured text data [41; 20; 4; 23]. The research around applying data mining techniques to unstructured text is termed knowledge discovery in texts (KDT) [45], or text data mining [62], or text mining [115]. Hence we could consider text mining as an instance of Web content mining. We discuss text mining further in the next section. We could differentiate the research done in Web content mining from two different points of view: IR and DB [30] views. The goal of Web content mining from the IR view is mainly to assist or to improve the information finding or filtering the information to the users usually based on either inferred or solicited user profiles, while the goal of Web content mining from the DB view mainly tries to model the data on the Web and to integrate them so that more sophisticated queries other than the keywords based search could be performed. These viewpoints are further discussed in the next section.

Web structure mining [24] 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 could be used to discover authority sites for the subjects (authorities) and overview sites for the subjects that point to many authorities (hubs).

Web usage mining [30] tries to make sense of the data generated by the Web surfer's sessions or behaviors. While the Web content and structure mining utilize the real or primary data on the Web, Web usage mining mines the secondary data derived from the interactions of the users while interacting with the Web. The Web usage data includes the data from Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data as the results of interactions. Table 1 gives an overview of the above Web mining categories (the explanations are given in the subsequent sections).

However we should emphasize that the distinctions between the above categories are not clear-cut. Web content mining might utilize text and links and even the profiles that are either inferred or inputted by the users. User profiles are mostly used for the user modeling applications or personal assistants. The same is true for Web structure mining that could use the information about the links in addition to the link structures. Moreover we could infer the traversed links from the documents that were requested during the user session from the logs generated by the server. We could also characterize the categories above from the point of view of the *scope* of most of the work done in the respective areas: local scope spans an individual Web site while global scope spans the entire Web. The scope of the Web content mining from the IR view and Web structure mining is global while the scope of the Web content mining from the DB view and Web usage mining is local. However this characterization is not clear-cut either.

In practice, the three Web mining tasks above could be used in isolation or combined in an application, especially in Web content and structure mining since the Web documents might also contain links. For example, Chakrabarti et al. [24] uses as Web content the terms in a document's link neighborhood and as Web structure the links from its neighbors, to classify Web pages. Joachims et al. [68] use Web content and usage to build a software tour agent for assisting users browsing a Web site.

2.3 Web Mining and the Agent Paradigm

Web mining is often viewed from or implemented within an agent paradigm. Thus, Web mining has a close relationship with software agents or intelligent agents. Indeed some of these agents perform data mining tasks to achieve their goals. According to Green, et al. [58] there are three sub-categories of software agents: user interface agents, distributed agents, and mobile agents. The sub-categories of software agents that are relevant for data mining tasks are user interface agents and distributed agents. User interface agents try to maximize the productivity of current users interaction with the system by adapting behavior. The issue of personalization abounds here. User interface agents that can be classified into the Web mining agent category are information retrieval agents, information filtering agents, and personal assistant agents. Distributed agents technology is concerned with problem solving by a group of agents and relevant agents in this category are distributed agents for knowledge discovery or data mining (for example see [71]).

Table 2: The association between the categories of Web mining and the agent paradigm

Content-based filters	↔	Content mining
Reputation-based filters	↔	Structure (and content) mining
Collaborative or social-based filters	↔	Usage mining
Event-based filters	↔	Usage mining
Hybrid filters	↔	Combination of the categories

There are two frequently used approaches for developing intelligent agents that help users find and retrieve relevant information from the Web [11], namely content-based and collaborative approaches. In the content-based approach, the system searches for items that match based on an analysis of the content using the user preferences. In the collaborative approach, the system tries to find users with similar interests to give recommendations to. The system does this by analyzing the user profiles and sessions or transactions. It assumes that if some users rate an item high, then the other users with similar interests would rate this item high also. So this approach mainly uses the usage data (user ratings). Viewed in this light we could categorize the content-based methods as Web content mining and categorize the collaborative approaches as Web usage mining. However, collaborative approaches might also be used or combined with the Web content.

A similar view related to the Web mining categories above also exists in the software agent community. Delgado [38] classifies the user interface agents by the underlying information filtering technology into content-based filters, reputation based filters, collaborative or social-based filters, event-based filters, and hybrid filters. In event based filtering, the system tracks and follows the events that are inferred from the surfing habits of people in the Web. Some examples of those events are saving a URL into a bookmark folder, mouse clicks and scrolls, link traverse behavior, etc. We could make an association between these agent-based categories with the Web mining categories above. Table 2 shows the association.

3. WEB CONTENT MINING

In this section we list some of the research in the respective categories in separate tables. We should note that the lists are by no means complete. The explanations on the methods surveyed are beyond the scope of this paper. Interested

Table 1: Web mining categories

	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR View	DB 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 patterns in text - User modeling	- Finding frequent substructures - Web site schema discovery	- Categorization - Clustering	- Site construction, adaptation, and management - Marketing - User modeling

readers can consult the book by Mitchell [89] and the respective papers for the explanation of the methods. We just intend to give a taste on the variety of some representations, processes, methods, and applications that have been used.

3.1 Information Retrieval View

3.1.1 *Information Retrieval View for Unstructured Documents*

Table 3 summarizes some of the research done for unstructured documents. What we mean by the unstructured documents is free texts such as news stories. Most of the research in table 3 uses bag of words to represent unstructured documents. The bag of words or vector representation [107] takes single words found in the training corpus as features. This representation ignores the sequence in which the words occur and is based on the statistic about single words in isolation. The features could be Boolean (a word either occurs or does not occur in a document), or frequency based (frequency of the word in a document). Variations of the feature selection include removing the case, punctuation, infrequent words, and stop words. The features could be reduced further by applying some other feature selection techniques, such as information gain, mutual information, cross entropy, or odds ratio (see [92] for the details). Other preprocessing includes Latent Semantic Indexing (LSI) [36] that seeks to transform the original document vectors to a lower dimensional space by analyzing the correlational structure of terms in the document collection such that similar documents that do not share terms are placed in the same topic, and stemming which reduces words to their morphological roots. For example the words “informing”, “information”, “informer”, and “informed” would be stemmed to their common root “inform” and only the latter word is used as the feature instead of the former four. While those pre-processing variations are useful for reducing feature set size, the generality of their effectiveness over different domains for text categorization tasks are doubted [106].

Other feature representations are also possible such as using information about word positions in the document [27; 3; 51], using n-grams representation (word sequences of length up to n) [65; 71] (for example “the morphological roots” is a tri-gram), using phrases [39; 51; 108; 125] such as “the quick brown fox that run away”, using document concept categories [45], using terms [46] such as “annual interest rate” or “Wall Street”, using hypernyms (linguistic term for the “is a” relationship - a dog is an animal, thus “animal” is a hypernym of “dog”) [108], or using named entities [124] such as people’s names, dates, email addresses, locations, organizations, or URLs. The relational representation ([27; 69] in table 3) that we mean here is actually first order logic, a language that is more expressive than propositional logic (for instance see [89]). For example in the bag of words representation features are the frequencies of specific words; using a relational representation one might use relationships between different words and their positions, e.g. “word X is to the left of word Y in the same sentence”. Although different types of representations have been used, there is currently no study that shows clear advantages of some representations over several domains for text categorization tasks [91]. Indeed, Scott and Matwin [108] compare different representations (bag of words, phrase based, and hypernym) but found no significant differences in the performance of different representations.

As we can see from table 3, the commonly used process is 1 - 2 - 3 - 4, while some others do not use any or only use a minimal pre-processing step (process 1 - 3 - 4). The name and explanation of the four steps are described in section 2.1 above. The use of text compression techniques [124] is rather new for the text classification task. The applications range from text classification or categorization, event detection and tracking, finding extraction patterns or rules, to finding some interesting patterns in the text documents. Event detection and tracking problems are sub-topics of a broader initiative called topic detection and tracking (TDT), which is a new line of research related to research in information retrieval and filtering [6]. TDT is an initiative to investigate the state of the art in finding and following new events in a stream of news stories broadcast [5].

Recently the usage of the term text mining has been a subject to controversy. There are at least two controversies that we are aware of: one is regarding the usage of the term “mining” itself [62] and the other one is regarding the meaning of the word “knowledge” in knowledge discovery from text (KDT) [75]. As far as we know, the term text mining or KDT was first proposed by Feldman and Dagan in [45]. They suggest to structure the text documents by means of information extraction, text categorization, or applying

Table 3: An IR view on Web content mining for unstructured documents

Author	Document Representation	Process	Method	Application
Ahonen, et al. [3]	Bag of words and word positions	1 - 2 - 3 - 4	Episode rules	- Finding keywords and keyphrases - Discovering grammatical rules and collocations
Billsus and Pazzani [14]	Bag of words	1 - 2 - 3 - 4	- TFIDF - Naïve Bayes	Text classification
Cohen [27]	Relational	1 - 2 - 3 - 4	- Propositional rule based system Inductive Logic Programming	Text classification
Dumais, et al. [39]	- Bag of words - Phrases	1 - 2 - 3 - 4	- TFIDF - Decision trees - Naive Bayes - Bayes nets - Support Vector Machines	Text categorization
Feldman and Dagan [45]	Concept categories	1 - 2 - 3 - 4	Relative entropy	Finding patterns between concept distributions in textual data
Feldman, et al. [46]	Terms	1 - 2 - 3 - 4	Association rules	Finding patterns across terms in textual data
Frank, et al. [51]	Phrases and their positions	1 - 2 - 3 - 4	Naive Bayes	Extracting keyphrases from text documents
Freitag and McCallum [53]	Bag of words	1 - 3 - 4	Hidden Markov Models	Learning extraction models
Hofmann [63]	Bag of words	1 - 2 - 3 - 4	Unsupervised statistical method	Hierarchical clustering
Honkela, et al. [65]	Bag of words with n-grams	1 - 2 - 3 - 4	Self-Organizing Maps	Text and document clustering
Junker, et al. [69]	Relational	1 - 2 - 3 - 4	Inductive Logic Programming	- Text categorization - Learning extraction rules
Kargupta, et al. [71]	Bag of words with n-grams	1 - 2 - 3 - 4	- Unsupervised hierarchical clustering - Decision trees - Statistical analysis	Text classification and hierarchical clustering
Nahm and Mooney [95]	Bag of words	1 - 2 - 3 - 4	Decision trees	Predicting (words) relationship
Nigam, et al. [98]	Bag of words	1 - 3 - 4	Maximum entropy	Text classification
Scott and Matwin [108]	- Bag of words - Phrases - Hypernyms and synonyms	1 - 2 - 3 - 4	Rule based system	Text classification
Soderland [111]	Sentences, and clauses	1 - 2 - 3 - 4	Rule learning	Learning extraction rules
Weiss, et al. [121]	Bag of words	1 - 2 - 3 - 4	Boosted decision trees	Text categorization
Wiener, et al. [122]	Bag of words	1 - 2 - 3 - 4	- Neural Networks - Logistic Regression	Text categorization
Witten, et al. [124]	Named entity	1 - 2 - 3 - 4	Text compression	Named entity classifier
Yang, et al. [125]	Bag of words and phrases	1 - 2 - 3 - 4	- Clustering algorithms - k-Nearest Neighbor - Decision tree	Event detection and tracking

NLP techniques as pre-processing step before performing any kind of KDTs. The reason is mining on the unprepared documents does not provide effectively exploitable results [103; 46]. Currently the term text mining has been used to describe different applications such as text categorization [64; 115; 121], text clustering [115; 104], IE [3], empirical computational linguistic tasks [62], exploratory data analysis [62], finding patterns in text databases [45; 46], finding sequential patterns in texts [83; 3; 4], and association discovery [115; 95]. So although some of the papers surveyed mention their application as text mining, we use less controversial names for the applications.

3.1.2 Information Retrieval View for Semi-Structured Documents

We can see from table 4 that the process used in the works surveyed above is 1 - 2 - 3 - 4. We can also see that the works surveyed in table 4 use richer representations compared to the works surveyed in table 3. This is due to the additional structural (HTML and hyperlink) information in the hypertext documents. Actually all of the works surveyed utilize the HTML structures inside the documents and some utilize the hyperlink structure between the documents for document representation. The methods that are used are common data mining methods. The applications ranged from hypertext classification or categorization and clustering, learning relations between Web documents, learning extraction patterns or rules, and finding patterns in semi-structured data.

3.2 Database View

As mentioned in [50], the database techniques on the Web are related to the problems of managing and querying the information on the Web. [50] mentions that there are three classes of tasks related to those problems: modeling and querying the Web, information extraction and integration, and Web site construction and restructuring. Although the first two tasks are related to Web content mining applications, not all the works there are inside the scope of Web content mining. This is due to the absence of the machine learning or data mining techniques in the process. 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 become possible. As mentioned previously, the DB view of Web content mining mainly tries to model the data on the Web and to integrate them so that more sophisticated queries other than the keywords based search could be performed. These could be achieved by finding the schema of Web documents, building a Web warehouse or a Web knowledge base or a virtual database. The research done in this area mainly deals with semi-structured data. Semi-structured data from database view often refers to data that has some structure but no rigid schema [1; 18].

From table 5, we can see that the DB view uses representations that differ from the IR view that we see in table 3 and table 4. The DB view mainly uses Object Exchange Model (OEM) [2] that represents semi-structured data by a labeled graph. The data in the OEM is viewed as a graph, with objects as the 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. All of the processes that are surveyed above are 1 - 2 - 3 - 4. However, the process used here typically starts from manually selected Web sites for doing Web content mining instead of searching the whole Internet for the specific resources. This is partly due to the applications of the DB view that are quite different from those of the IR view (which mostly are classification tasks). The process 1 and 2 is typically done by site-specific wrappers or parsers for hypertext documents.

Most of the applications that are surveyed above are the task of schema extraction or discovery [70; 116] or building DataGuides [56; 96; 57]. Roughly speaking, a schema or DataGuide is a kind of structural summary of semi-structured data. For practical applications and computational reason, this summary is often approximated [1; 57]. Some applications do not deal with the task of finding the global schema but deal with the task of finding frequent substructures (subschema) in semi-structured data. Another application deals with the task of creating multi-layered database (MLDB) [127] in which each layer is obtained by generalizations on lower layers and use a special purpose query language for Web mining to extract some knowledge from the MLDB of Web documents. This is an example of the query perspective of data mining. There has been some work on query languages for semi-structured data [2; 19] and for the Web [8; 79; 82; 48]. However, we only see the works by Zaïane, et al. [127] and Singh, et al. [109] that are inside the scope of Web content mining.

Due to the different representations used in the DB view, most of the methods used for data mining are also different except the ILP methods that could operate on relational or graphical data. These differences are partly due to the inappropriateness of many existing data mining techniques, which operate on flat data, to operate on relational or graphical data. [59; 96; 127] use proprietary algorithms for schema discovery and for the construction of MLDB, [70] uses a modified version of association rules, and [116] uses an upgraded first order logic version of association rules [37].

3.3 About Mining Multimedia Data

We should note that we have not actually discussed the issue of mining multimedia data on the Web. Although multimedia data has been the major focus for many researchers [73; 114] and many techniques for multimedia IR and extraction have been proposed (for example see [61]), multimedia data mining is still in its infancy [128]. Uthurusamy [117], Shapiro et al. [102], and Mitchell [90] assert that working towards a unifying framework for representation, problem solving, and learning from multimedia data is indeed a challenge. Fayyad et al. [42] describes mining the images of sky objects taken from satellite. Smyth, et al. [110] describes mining images to identify small volcanoes on Venus. More recent works are [128] in the application of Web data warehousing and [66] in the application of a medical IR system for mining the multimedia data on the Web. For a definition and a short survey on multimedia data mining, we refer to [128].

4. WEB STRUCTURE MINING

If in the database view of Web content mining we are interested in the structure within Web documents (intra-document structure), in Web structure mining we are interested in the structure of the hyperlinks within the Web itself (inter-

Table 4: An IR view on Web content mining for semi-structured documents

Author	Document Representation	Process	Method	Application
Craven, et al. [34]	Relational and ontology	1 - 2 - 3 - 4	- Modified Naive Bayes - Inductive Logic Programming	- Hypertext classification - Learning Web page relation - Learning extraction rules
Crimmins, et al. [35]	Phrase, URLs, and meta information	1 - 2 - 3 - 4	Unsupervised and supervised classification algorithms	- Hierarchical and graphical classification - Clustering
Fürnkranz [54]	Bag of words and hyperlinks information	1 - 2 - 3 - 4	Rule learning	Hypertext classification
Joachims, et al. [68]	Bag of words and hyperlinks information	1 - 2 - 3 - 4	- TFIDF - Reinforcement learning	Hypertext prediction
Muslea, et al. [94]	Bag of words, tags, and word positions	1 - 2 - 3 - 4	Rule learning	Learning extraction rules
Shavlik and Eliassi-Rad [40]	Localized bag of words, and relational.	1 - 2 - 3 - 4	Neural networks with reinforcement learning	Hypertext (homepage) classification
Singh, et al. [109]	Concepts and Named entity	1 - 2 - 3 - 4	- Modified association rule - Classification algorithm	Finding patterns in semi-structured texts
Soderland [111]	Sentences, phrases, and named entity	1 - 2 - 3 - 4	Rule learning	Learning extraction rules

Table 5: Web content mining from a database view

Author	Document Representation	Process	Method	Application
Goldman and Widom [57]	OEM	1 - 2 - 3 - 4	Proprietary algorithms	Finding DataGuide in semi-structured data
Grumbach and Mecca [59]	Strings and relational	1 - 2 - 3 - 4	Proprietary algorithms	Finding schema in semi-structured data
Nestorov, et al. [96]	OEM	1 - 2 - 3 - 4	Proprietary algorithms	Finding type hierarchy in semi-structured data
Toivonen [116]	OEM	1 - 2 - 3 - 4	Upgraded association rules	Finding useful sub-structure in semi-structured data
Wang and Liu [70]	OEM	1 - 2 - 3 - 4	Modified association rules	Finding frequent sub-structures in semi-structured data
Zaiane and Han [127]	Relational	1 - 2 - 3 - 4	Attribute-oriented induction	Multilevel databases

document structure). This line of research is inspired by the study of social networks and citation analysis [72; 23]. With social network analysis we could discover specific types of pages (such as hubs, authorities, etc.) based on the incoming and outgoing links. Web structure mining utilizes the hyperlinks structure of the Web to apply social network analysis to model the underlying links structure of the Web itself. Research done by Kautz et al. [72] utilizes the network analysis of people to model the network of AI researchers. They use the name entity data found in close proximity in any public Web pages such as the hyperlinks from home pages, co-authorship and citation of papers, exchange of information between individuals found in net-news archives, and organization charts. In our framework, their research could be classified as a combination of Web structure and content mining.

Some algorithms have been proposed to model the Web topology such as HITS [74], PageRank [17] and improvements of HITS by adding content information to the links structure [24] and by using outlier filtering [13]. These models are mainly applied as a method to calculate the quality rank or relevancy of each Web page. Some examples are the Clever system [24] and Google [17]. Some other applications of the models include Web pages categorization [25] and discovering micro communities on the Web [76].

More applications of Web structure mining in the context of Web warehouse are discussed by Madria, et al. [85]. These include measuring the completeness of the Web sites by measuring the frequency of local links that reside in the same server, measuring the replication of Web documents across the Web warehouse that helps in identifying the mirrored sites for example, 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.

5. WEB USAGE MINING

Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web. As mentioned before, the mined data in this category are the secondary data on the Web as the result of interactions. These data could range very widely but generally we could classify them into the usage data that reside in the Web clients, proxy servers and servers [113]. The Web usage mining process could be classified into two commonly used approaches [16]. The first approach maps the usage data of the Web server into relational tables before an adapted data mining technique is performed. The second approach uses the log data directly by utilizing special pre-processing techniques. As is true for typical data mining applications, the issues of data quality and pre-processing are also very important here. The typical problem is distinguishing among unique users, server sessions, episodes, etc. in the presence of caching and proxy servers [87; 113]. For the details and comparison of some pre-processing methods for Web usage data we refer to [31].

In general, typical data mining methods (see for example in [113]) 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 composite association rules [15], an extension of a traditional sequence discovery algorithm (MIDAS [12]), and hypertext probabilistic grammars [16]. The Web usage data could also be represented with graphs [12; 99]. Often the Web usage mining uses some background or domain knowledge such as navigation templates, Web content, site topology, concept hierarchies, and syntactic constraints [12; 112].

The applications of Web usage mining could be classified into two main categories: learning a user profile or user modeling in adaptive interfaces (personalized) (for examples see [80]) and learning user navigation patterns (impersonalized) (for examples see [112]). Web users would be interested in, among others, techniques that could learn their information needs and preferences, which is user modeling possibly combined with Web content mining. On the other hand, information providers would be interested in, among others, techniques that could improve the effectiveness of the information on their Web sites by adapting the Web site design or by biasing the user's behavior towards satisfying the goals of the site. In other words, they are interested in learning user navigation patterns. Then the learned knowledge could be used for applications such as personalization (at a Web site level), system improvement, site modification, business intelligence, and usage characterization (see [113] for the detail). It is not in our intention to give a complete survey of Web usage mining research here. Interested readers could consult the overview papers by Srivastava, et al. [113], Spiliopoulou [112], and Masand and Spiliopoulou [87], and Robert Cooley's Ph.D. thesis [32] for mining user patterns and the overview paper by Langley [80] for mining user profiles.

6. RELATED WORKS

As far as we know, it was Etzioni [41] who first coined the term Web mining. Etzioni starts by making a hypothesis that the information on the Web is sufficiently structured and outlines the subtasks of Web mining. His paper describes the Web mining processes. There have been some works around the survey of data mining on the Web. The first paper that we know that noticed the confusion in the Web mining research is [30]. It gives a Web mining taxonomy but restricted to Web content and Web usage mining, and gives a survey on Web usage mining. It divides the Web content mining into the agent based approach and the database approach. We use a similar division but divide it into the IR approach instead of the agent approach. Later, in [113] they classify Web mining into three categories that are similar to our categories. Compared to their paper, our paper points out three confusions on the usage of the term Web mining, identifies additional user-centered Web mining processes, and provides new perspectives for the Web mining categories. We use the Web mining categories suggested in [85] and [16]. [16] proposes a new model for mining Web log data, while [85] discusses the research issues of Web mining in the context of Web warehouse project.

Carbonell et al. [20] give an overview of the workshop on learning from text and the Web that is related to Web content (from the IR view) and usage mining. They also give an outline of the research directions in that area. Mladenic [91] surveys the research on text learning and related intelligent agents. She compares two frequently used approaches for developing intelligent agents, namely collaborative and content based. In our categories, these would be Web content (from the IR view) and usage mining. She also surveys research on machine learning applied to text data, which is broader than but similar to our discussion in section 3.1.1 about the IR view of Web content mining from unstructured documents. Carbonell et al. [21] review the emerging research collaborations between the IR and machine learning communities in a special issue of the Machine Learning journal. They also indicate some fertile research areas for both communities. Garofalakis et al. [55] review some data mining techniques and the algorithms for Web mining that specifically take into account the hyperlink information. Chakrabarti [23] provides a survey of data mining for hypertext. His paper mainly surveys the statistical techniques for Web content across the continuum of supervised, semi-supervised and unsupervised learning, and social network analysis techniques for Web structure mining. Levy and Weld [84] wrote a survey in the special issue of Artificial Intelligence on intelligent Internet systems that we think describes a broader domain than Web mining. Vaithyanathan [118] gives an overview of the papers in the special issue of Artificial Intelligence Review on data mining on the Internet. He mentions similar categories of Web mining as ours, except the database view of Web content mining. Some other related work that we found recently in special issues of some magazines are the following. Yang and Pedersen edited a special issue on intelligent information retrieval [126]. Filman and Pant edited a special issue on searching the Internet [49].

7. CONCLUSIONS

In this paper we survey the research in the area of Web mining. We point out some confusions regarded the usage of the term Web mining. We also suggest three Web mining categories and then situate some of the research with respect to these categories. We also explore the connection between Web mining categories and the related agent paradigm. For the survey, we focus on representation issues, on the process, and on the learning algorithm, and the application of the recent works as the criteria. The Web presents new challenges to the traditional data mining algorithms that work on flat data. We have seen that some of the traditional data mining algorithms have been extended or new algorithms have been used to work on the Web data.

An interesting direction of Web content mining is the recent interest in information integration [26; 47], which could be in the form of a Web knowledge base [20; 29] or Web warehouse [85], or in the form of a mediator (see [47] for examples). At least this is the area where database and other research communities such as IR, AI, and machine learning met recently. Information integration was mainly concerned with integrating various databases but has changed its focus with the increasing popularity of the Web [47]. The same is also true for the research in IE, which could be thought as a mediator or wrapper in the information integration area. Information integration also raises some other research questions such as scaling up the number of Web sites that could be integrated, wrapper maintenance, building and maintaining a global schema, etc. [28] (see also [77] for other issues).

Topic detection and tracking (TDT) is also a promising research area for IR and machine learning communities that raises, among others, temporal issue in the data. It would be interesting if the learning algorithm could model this aspect accurately. Some other promising research issues in the area of Web content mining are discussed in [20]. Finally, another 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 this representation or that could take advantage of the available structures on the Web.

8. ACKNOWLEDGEMENTS

We thank Johannes Fürnkranz, Dunja Mladenic, Nico Jacobs, and Maurice Bruynooghe for reading the draft and the useful comments. We also thank Luc Dehaspe for the interesting discussions. Raymond Kosala is supported by the GOA Project LP+ of the Katholieke Universiteit Leuven. Hendrik Blockeel is a post-doctoral fellow of the Fund for Scientific Research of Flanders.

9. REFERENCES

[1] S. Abiteboul. Querying semi-structured data. In F. N. Afrati and P. Kolaitis, editors, *Database Theory - ICDT '97, 6th International Conference, Delphi, Greece, January 8-10, 1997, Proceedings*, volume 1186 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 1997.

[2] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. L. Wiener. The lorel query language for semistructured data. *Int. J. on Digital Libraries*, 1(1):68–88, 1997.

[3] H. Ahonen, O. Heinonen, M. Klemettinen, and A. Verkamo. Applying data mining techniques for descriptive phrase extraction in digital document collections. In *Advances in Digital Libraries (ADL'98), Santa Barbara, California, USA, April 1998*, 1998.

[4] H. Ahonen, O. Heinonen, M. Klemettinen, and A. Verkamo. Finding co-occurring text phrases by combining sequence and frequent set discovery. In R. Feldman, editor, *Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, pages 1–9, 1999.

[5] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998*, 1998.

[6] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval August 24 - 28, 1998*, pages 37–45, Melbourne Australia, 1998.

[7] D. E. Appelt and D. Israel. Introduction to information extraction technology. In *Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99, Tutorial*, 1999.

[8] G. O. Arocena and A. O. Mendelzon. Weboql: Restructuring documents, databases, and webs. *Theory and Practice of Object Systems*, 5(3):127–141, 1999.

[9] P. Atzeni and G. Mecca. Cut & paste. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 12-14, 1997, Tucson, Arizona*, pages 144–153. ACM Press, 1997.

[10] R. Baeza-Yates and e. Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Company, 1999.

[11] M. Balabanovi'c and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–70, 1997.

[12] A. Büchner, M. Baumgarten, S. Anand, M. Mulvenna, and J. Hughes. Navigation pattern discovery from internet data. In *Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999*, San Diego, CA, USA, 1999.

[13] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval August 24 - 28, 1998*, pages 104–111, Melbourne Australia, 1998.

[14] D. Billsus and M. Pazzani. A hybrid user model for news story classification. In *Proceedings of the Seventh International Conference on User Modeling (UM '99)*, Banff, Canada, 1999.

[15] J. Borges and M. Levene. Mining association rules in hypertext databases. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), August 27-31, 1998, New York City, New York, USA*, 1998.

[16] J. Borges and M. Levene. Data mining of user navigation patterns. In *Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999, San Diego, CA, USA*, pages 31–36, 1999.

[17] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Seventh International World Wide Web Conference*, Brisbane, Australia, 1998.

[18] P. Buneman. Semistructured data. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 12-14, 1997, Tucson, Arizona*, pages 117–121. ACM Press, 1997.

[19] P. Buneman, S. B. Davidson, G. G. Hillebrand, and D. Suciu. A query language and optimization techniques for unstructured data. In H. V. Jagadish and I. S. Mumick, editors, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, June 4-6, 1996*, pages 505–516. ACM Press, 1996.

[20] J. Carbonell, M. Craven, S. Fienberg, T. Mitchell, and Y. Yang. Report on the conald workshop on learning from text and the web. In *CONALD Workshop on Learning from Text and the Web, June, 1998*, 1998.

[21] J. Carbonell, Y. Yang, and W. Cohen. Special issue of machine learning on information retrieval introduction. *Machine Learning*, 39:99–101, 2000.

[22] C. Cardie. Empirical methods in information extraction. *AI Magazine*, 18(4):65–79, 1997.

[23] S. Chakrabarti. Data mining for hypertext: A tutorial survey. *ACM SIGKDD Explorations*, 1(2):1–11, 2000.

[24] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Mining the link structure of the world wide web. *IEEE Computer*, 32(8):60–67, 1999.

[25] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In L. M. Haas and A. Tiwary, editors, *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 307–318. ACM Press, 1998.

[26] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The tsimmis project: Integration of heterogeneous information sources. In *Proceedings of the 10th Meeting of the Information Processing Society of Japan*, pages 7–18, 1994.

[27] W. W. Cohen. Learning to classify english text with ilp methods. In *Advances in Inductive Logic Programming (Ed. L. De Raedt), IOS Press*, 1995.

[28] W. W. Cohen. Some practical observations on integration of web information. In *ACM SIGMOD Workshop on The Web and Databases (WebDB'99)*, pages 55–60, Philadelphia, Pennsylvania, USA, 1999.

[29] W. W. Cohen. What can we learn from the web? In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML'99)*, pages 515–521, 1999.

[30] R. Cooley, B. Mobasher, and 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)*, 1997.

[31] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1), 1999.

[32] R. W. Cooley. *Web Usage Mining: Discovery and Application of Interesting Patterns from Web data*. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000.

[33] J. Cowie and W. Lehnert. Information extraction. *Communications of the ACM*, 39(1):80–91, 1996.

[34] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the world wide web. In *Proceedings of the Fifteenth National Conference on Artificial Intellligence (AAAI98)*, pages 509–516, 1998.

[35] F. Crimmins, A. Smeaton, T. Dkaki, and J. Mothe. Tétrafusion: Information discovery on the internet. *IEEE Intelligent Systems*, 14(4):55–62, 1999.

[36] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[37] L. Dehaspe and L. de Raedt. Mining association rules in multiple relations. In *Proceedings of the 7th International Workshop on Inductive Logic Programming*, volume 1297 of *Lecture Notes in Computer Science*, pages 125–132, Prague, Czech Republic, 1997. Springer.

[38] J. A. Delgado. *Agent-Based Information Filtering and Recommender System On the Internet*. PhD thesis, Dept. of Intelligence Computer Science, Nagoya Institute of Technology, March 2000.

[39] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 1998 ACM 7th international conference on Information and knowledge management*, pages 148–155, Washington United States, 1998.

[40] J. S. . T. Eliassi-Rad. Intelligent agents for web-based tasks: An advice-taking approach. In *Working Notes of the AAAI/ICML-98 Workshop on Learning for Text Categorization, Madison, WI*, pages 588–589, 1999.

[41] O. Etzioni. The world wide web: Quagmire or gold mine. *Communications of the ACM*, 39(11):65–68, 1996.

[42] U. Fayyad, S. Djorgovski, and N. Weir. Automating the analysis and cataloging of sky surveys. In *Advances in Knowledge Discovery and Data Mining*, pages 471–493. AAAI Press, 1996.

[43] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press, 1996.

[44] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: toward a unifying framework. In *Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining*, pages 82–88, 1996.

[45] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pages 112–117, Montreal, Canada, 1995.

[46] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir. Text mining at the term level. In *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98*, volume 1510 of *Lecture Notes in Computer Science*, pages 56–64. Springer, 1998.

[47] D. Fensel, C. Knoblock, N. Kushmerick, and M.-C. Rousset. Workshop on intelligent information integration (iii'99). *AI Magazine*, 21(1):91–94, 2000.

[48] M. F. Fernandez, D. Florescu, A. Y. Levy, and D. Suciu. A query language for a web-site management system. *SIGMOD Record*, 26(3):4–11, 1997.

[49] R. E. Filman and S. Pant. Searching the internet - guest editors' introduction. *IEEE Internet Computing*, 2(4):21–23, 1998.

[50] D. Florescu, A. Y. Levy, and A. O. Mendelzon. Database techniques for the world-wide web: A survey. *SIGMOD Record*, 27(3):59–74, 1998.

[51] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. Domain-specific keyphrase extraction. In *Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99*, pages 668–673, 1999.

[52] D. Freitag. Information extraction from html: Application of a general learning approach. In *Proceedings of the Fifteenth Conference on Artificial Intelligence AAAI-98 (1998)*, pages 517–523, 1998.

[53] D. Freitag and A. McCallum. Information extraction with hmms and shrinkage. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

[54] J. Fürnkranz. Exploiting structural information for text classification on the www. In *Advances in Intelligent Data Analysis, Third International Symposium, IDA-99*, pages 487–498, 1999.

[55] M. N. Garofalakis, R. Rastogi, S. Seshadri, and K. Shim. Data mining and the web: Past, present and future. In *Workshop on Web Information and Data Management, 1999*, pages 43–47, 1999.

[56] R. Goldman and J. Widom. Dataguides: Enabling query formulation and optimization in semistructured databases. In M. Jarke, M. J. Carey, K. R. Dittrich, F. H. Lochovsky, P. Loucopoulos, and M. A. Jeusfeld, editors, *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pages 436–445. Morgan Kaufmann, 1997.

[57] R. Goldman and J. Widom. Approximate dataguides. In *Proceedings of the Workshop on Query Processing for Semistructured Data and Non-Standard Data Formats*, 1999.

[58] S. Green, L. Hurst, B. Nangle, P. Cunningham, F. Somers, and R. Evans. Software agents: A review. Technical Report TCD-CS-1997-06, Technical Report of Trinity College, University of Dublin, 1997.

[59] S. Grumbach and G. Mecca. In search of the lost schema. In *Database Theory - ICDT '99, 7th International Conference*, pages 314–331, 1999.

[60] J. Hammer, H. Garcia-Molina, J. Cho, A. Crespo, and R. Aranha. Extracting semistructured information from the web. In *Proceedings of the Workshop on Management of Semistructured Data*, pages 18–25, 1997.

[61] A. Hauptmann. Integrating and using large databases of text, image, video and audio. *IEEE Intelligent Systems*, 14(5):34–35, 1999.

[62] M. A. Hearst. Untangling text data mining. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[63] T. Hofmann. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In *Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99*, pages 682–687, 1999.

[64] S. J. Hong and S. M. Weiss. Advances in predictive model generation for data mining. Technical Report Report RC-21570, IBM Research Report, 1999.

[65] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Websom - self-organizing maps of document collections. In *Proc. of Workshop on Self-Organizing Maps 1997 (WSOM'97)*, pages 310–315, 1997.

[66] A. Houston, H. Chen, S. M. Hubbard, B. R. Schatz, T. D. Ng, R. R. Sewell, and K. M. Tolle. Medical data mining on the internet: Research on a cancer information system. *Artificial Intelligence Review*, 13:437–446, 1999.

[67] C.-N. Hsu and M.-T. Dung. Generating finite-state transducers for semi-structured data extraction from the web. *Information Systems*, 23(8):521–538, 1998.

[68] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-97*, pages 770–777, 1997.

[69] M. Junker, M. Sintek, and M. Rinck. Learning for text categorization and information extraction with ilp. In *Proceedings of the Workshop on Learning Language in Logic, Bled, Slovenia, 1999*, 1999.

[70] H. L. K. Wang. Discovering association of structure from semistructured objects. *To appear in IEEE Transactions on Knowledge and Data Engineering*, 1999.

[71] H. Kargupta, I. Hamzaoglu, and B. Stafford. Distributed data mining using an agent based architecture. In *Proceedings of Knowledge Discovery And Data Mining*, pages 211–214. AAAI Press, 1997.

[72] H. Kautz, B. Selman, and M. Shah. The hidden web. *AI magazine*, 18(2):27–36, 1997.

[73] S. Khoshafian and A. B. Baker. *Multimedia and Imaging Databases*. Morgan Kaufmann Publishers, 1996.

[74] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. of ACM-SIAM Symposium on Discrete Algorithms, 1998*, pages 668–677, 1998.

[75] Y. Kodratoff. About knowledge discovery in texts: A definition and an example. In *Proc. of Advanced Course on Artificial Intelligence 1999 (ACAI-99) on Machine Learning Applications (Invited talk)*, 1999.

[76] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proceedings of the Eighth World Wide Web Conference (WWW8)*, 1999.

[77] N. Kushmerick. Gleaning the web. *IEEE Intelligent Systems*, 14(2):20–22, 1999.

[78] N. Kushmerick, D. Weld, and R. Doorenbos. Wrapper induction for information extraction. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-97*, pages 729–737, 1997.

[79] L. Lakshmanan, F. Sadri, and I. Subramanian. A declarative language for querying and restructuring the web. In *Proceedings of 6th. International Workshop on Research Issues in Data Engineering, RIDE '96*, pages 12–21, 1996.

[80] P. Langley. User modeling in adaptive interfaces. In *Proceedings of the Seventh International Conference on User Modeling*, pages 357–370, 1999.

[81] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999.

[82] lberto O. Mendelzon, G. A. Mihaila, and T. Milo. Querying the world wide web. In *Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems*, pages 80–91, 1996.

[83] B. Lent, R. Agrawal, and R. Srikant. Discovering trends in text databases. In *Proc. 3 rd Int Conf. On Knowledge Discovery and Data Mining (KDD 1997)*, pages 227–230, 1997.

[84] A. Y. Levy and D. S. Weld. Intelligent internet systems. *Artificial Intelligence*, 118(1-2), 2000.

[85] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim. Research issues in web data mining. In *Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99*, pages 303–312, 1999.

[86] P. Maes. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, 1994.

[87] B. Masand and M. Spiliopoulou. Webkdd-99: Workshop on web usage analysis and user profiling. *SIGKDD Explorations*, 1(2), 2000.

[88] A. McCallum, K. Nigam, J. Rennie, and K. Seymore. A machine learning approach to building domain-specific search engines. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-99*, pages 662–667, 1999.

[89] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[90] T. M. Mitchell. Machine learning and data mining. *Communications of the ACM*, 42(11):30–36, 1999.

[91] D. Mladenic. Text-learning and related intelligent agents. *IEEE Intelligent Systems*, 14(4):44–54, 1999.

[92] D. Mladenic and M. Grobelnik. Feature selection for unbalanced class distribution and naïve bayes. In *Proceedings of the 16th International Conference on Machine Learning ICML-99*, pages 258–267, 1999.

[93] I. Muslea. Extraction patterns for information extraction tasks: A survey. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, 1999.

[94] I. Muslea, S. Minton, and C. Knoblock. Wrapper induction for semistructured, web-based information sources. In *Proceedings of the Conference on Automatic Learning and Discovery CONALD-98*, 1998.

[95] U. Y. Nahm and R. J. Mooney. Ua mutually beneficial integration of data mining and information extraction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-00)*, 2000.

[96] S. Nestorov, S. Abiteboul, and R. Motwani. Infering structure in semistructured data. *SIGMOD Record*, 26(4), 1997.

[97] S. Nestorov, S. Abiteboul, and R. Motwani. Extracting schema from semistructured data. In L. M. Haas and A. Tiwary, editors, *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 295–306. ACM Press, 1998.

[98] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[99] G. Paliouras, C. Papatheodorou, V. Karkaletsis, P. Tzitziras, and C. D. Spyropoulos. Large-scale mining of usage data on web sites. In *AAAI 2000 Spring Symposium on Adaptive User Interfaces*, 2000.

[100] M. T. Pazienza, editor. *Information Extraction: A multidisciplinary Approach to an Emerging Information Technology*, volume 1299 of *Lecture Notes in Computer Science*. International Summer School, SCIE-97, Frascati (Rome), Springer, 1997.

[101] M. T. Pazienza, editor. *Information Extraction*, Frascati (Rome), 1999. International Summer School, SCIE-99 , Frascati (Rome).

[102] G. Piatetsky-Shapiro, R. Braachman, T. Khabaza, W. Kloesgen, and E. Simoudis. An overview of issues in developing industrial data mining and knowledge discovery applications. In *Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining, 1996*, pages 89–95, 1996.

[103] M. Rajman and R. Besançon. Text mining - knowledge extraction from unstructured textual data. In *Proc. of 6th Conference of International Federation of Classification Societies (IFCS-98), Roma (Italy)*, pages 473–480, 1998.

[104] A. Rauber and D. Merkl. Automatic labeling of self-organizing maps: Making a treasure-map reveal its secrets. In *Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining (PAKDD'99), Beijing, China*, 1999.

[105] J. Rennie and A. McCallum. Using reinforcement learning to spider the web efficiently. In *Proceedings of the 16th International Conference on Machine Learning ICML-99*, 1999.

[106] E. Riloff. Little words can make a big difference for text classification. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *SIGIR'95, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA, July 9-13, 1995 (Special Issue of the SIGIR Forum)*, pages 130–136. ACM Press, 1995.

[107] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill, 1983.

[108] S. Scott and S. Matwin. Feature engineering for text classification. In *Proceedings of the 16th International Conference on Machine Learning ICML-99*, 1999.

[109] L. Singh, B. Chen, R. Haight, P. Scheuermann, and K. Aoki. A robust system architecture for mining semi-structured data. In *Proceeding of The Second Int. Conference on Knowledge Discovery and Data Mining, 1998*, pages 329–333, 1998.

[110] P. Smyth, U. M. Fayyad, M. C. Burl, and P. Perona. Modeling subjective uncertainty in image annotation. *Advances in Knowledge Discovery and Data Mining*, pages 517–539, 1996.

[111] S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272, 1996.

[112] M. Spiliopoulou. Data mining for the web. In *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '99*, pages 588–589, 1999.

[113] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2), 2000.

[114] V. S. Subrahmanian. *Principles of Multimedia Database Systems*. Morgan Kaufmann Publishers, 1998.

[115] A.-H. Tan. Text mining: The state of the art and the challenges. In *Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases*, pages 65–70, 1999.

[116] H. Toivonen. On knowledge discovery in graph-structured data. In *Workshop on Knowledge Discovery from Advanced Databases (KDAD'99)*, pages 26–31, 1999.

[117] R. Uthurusamy. From data mining to knowledge discovery: Current challenges and future directions. In *Advances in Knowledge Discovery and Data Mining*, pages 561–569, 1996.

[118] S. Vaithyanathan. Introduction: Data mining on the internet. *Artificial Intelligence Review*, 13(5/6):343–344, 1999.

[119] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[120] K. Wang and H. Liu. Schema discovery for semistructured data. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, pages 271–274, 1997.

[121] S. M. Weiss, C. Apté, F. Damerau, D. E. Johnson, F. J. Oles, T. Goetz, and T. Hampp. Maximizing text-mining performance. *IEEE Intelligent Systems*, 14(4):63–69, 1999.

[122] W. Wiener, J. Pedersen, and A. Weigend. A neural network approach to topic spotting. In *Proceedings of the 4th Symposium on Document Analysis and Information Retrieval (SDAIR 95)*, pages 317–332, 1995.

[123] Y. Wilks. *Information Extraction as a core language technology*, volume 1299 of *Lecture Notes in Computer Science*, chapter In M-T. Pazienza (ed.), Information Extraction, pages 1–9. Springer, 1997.

[124] I. H. Witten, Z. Bray, M. Mahoui, and W. J. Teahan. Text mining: A new frontier for lossless compression. In *Data Compression Conference 1999*, pages 198–207, 1999.

[125] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, 1999.

[126] Y. Yang and J. Pedersen. Guest editors' introduction: Intelligent information retrieval. *IEEE Intelligent Systems*, 14(4):30–31, 1999.

[127] O. Zaïane and J. Han. Webml: Querying the world-wide web for resources and knowledge. In *Proc. ACM CIKM'98 Workshop on Web Information and Data Management (WIDM'98)*, pages 9–12, 1998.

[128] O. R. Zaiane, J. Han, Z.-N. Li, S. H. Chee, and J. Chiang. Multimediaminer: a system prototype for multimedia data mining. In *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pages 581–583, 1998.