Web Mining - Concepts, Applications and Research Directions

Jaideep Srivastava, Prasanna Desikan and Vipin Kumar

Department of Computer Science,
University of Minnesota
{srivasta, desikan, kumar}@cs.umn.edu

Abstract:
From its very beginning, the potential of extracting valuable knowledge from the Web has been quite evident. Web mining, i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage, is the collection of technologies to fulfill this potential. Interest in Web mining has grown rapidly in its short history, both in the research and practitioner communities. This paper provides a brief overview of the accomplishments of the field, both in terms of technologies and applications, and outlines key future research directions.

1 INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. A panel organized at ICTAI 1997 [91] asked the question "Is there anything distinct about Web mining (compared to data mining in general)?" While no definitive conclusions were reached then, the tremendous attention on Web mining in the past five years, and a number of significant ideas that have been developed, have answered this question in the affirmative in a big way. In addition, a fairly stable community of researchers interested in the area has been formed, largely through the successful series of WebKDD workshops, which have been held annually in conjunction with the ACM SIGKDD Conference since 1999 [53, 63, 64, 81], and the Web Analytics workshops, which have been held in conjunction with the SIAM data mining conference [46, 47]. A good survey of the research in the field till the end of 1999 is provided in [82] and [62].

Two different approaches were taken in initially defining Web mining. First was a 'process-centric view', which defined Web mining as a sequence of tasks [71]. Second was a 'data-centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process [29]. The

second definition has become more acceptable, as is evident from the approach adopted in most recent papers [45, 62, 82] that have addressed the issue. In this paper we follow the data-centric view of Web mining which is defined as,

> **Web mining** is the application of data mining techniques to extract knowledge from Web data, i.e. Web Content, Web Structure and Web Usage data.

The attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences. It is our attempt in this paper to capture them in a systematic manner, and identify directions for future research.

The rest of this paper is organized as follows : In Section 2 we provide a taxonomy of Web mining, in Section 3 we summarize some of the key concepts in the field, and in Section 4we describe successful applications of Web mining techniques. In 5 we present some directions for future research, and in Section 6 we conclude the paper.

2 WEB MINING TAXONOMY

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. We provide a brief overview of the three categories and a figure depicting the taxonomy is shown in Figure 1:

1. **Web Content Mining:** Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns , clustering of web documents and classification of Web Pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images, in the fields of image processing and computer vision, the application of these techniques to Web content mining has been limited.
2. **Web Structure Mining:** The structure of a typical Web graph consists of Web pages as nodes , and hyperlinks as edges connecting related pages. Web Structure Mining is the process of discovering structure information from the Web. This can be further divided into two kinds based on the kind of structure information used.
   - *Hyperlinks*: A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different

part of the same page is called an *Intra-Document Hyperlink*, and a hyperlink that connects two different pages is called an *Inter-Document Hyperlink*. There has been a significant body of work on hyperlink analysis, of which [75] provides an up-to-date survey.

- *Document Structure*: In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents [22, 55].

3. **Web Usage Mining:** Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [49]. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:
   - Web Server Data: The user logs are collected by Web server. Typical data includes IP address, page reference and access time.
   - Application Server Data: Commercial application servers, e.g. Weblogic [6], [11], StoryServer [94], etc. have significant features in the framework to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
   - Application Level Data: Finally, new kinds of events can always be defined in an application, and logging can be turned on for them - generating histories of these specially defined events.

   The usage data can also be split into three different kinds on the basis of the source of its collection: on the server side, the client side, and the proxy side. The key issue is that on the server side there is an aggregate picture of the usage of a service by all users, while on the client side there is complete picture of usage of all services by a particular client, with the proxy side being somewhere in the middle [49].

3 KEY CONCEPTS

In this section we briefly describe the key new concepts introduced by the Web mining research community.

3.1 Ranking metrics - for page quality and relevance.

Searching the Web involves two main steps: *Extracting the relevant pages to a query* and *ranking them according to their quality*. Ranking is important as it helps the user look for "quality" pages that are relevant to the query. Different metrics have been proposed to rank Web pages according to their quality. We briefly discuss two of the prominent metrics.

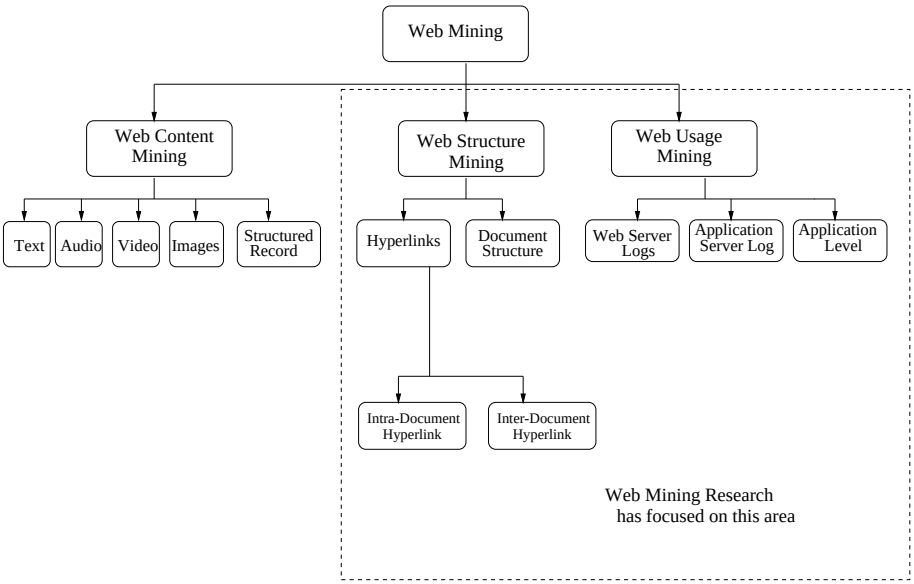

Fig. 1. Web mining Taxonomy

- **PageRank**: PageRank is a metric for ranking hypertext documents that determines the quality of these documents. Page et al. [61]developed this metric for the popular search engine Google [41, 83]. The key idea is that a page has a high rank if it is pointed to by many highly ranked pages. So, the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively till the rank of all the pages is determined. The rank of a page p can thus be written as:

$$PR(p) = d/n + (1-d) \sum_{(q,p) \in G} (\frac{PR(q)}{Outdegree(q)})$$

Here, n is the number of nodes in the graph and $OutDegree(q)$ is the number of hyperlinks on page q. Intuitively, the approach can be viewed as a stochastic analysis of a random walk on the Web graph. The first term in the right hand side of the equation corresponds to the probability that a random Web surfer arrives at a page p by typing the URL or from a bookmark, or may have a particular page as his/her homepage. Here ,d is the probability that a random surfer chooses a URL directly, rather than traversing a link[1] and $1-d$ is the probability that a person arrives at a page by traversing a link.. The second term in the right hand side of the equation corresponds to the probability of arriving at a page by traversing

[1] The parameter d, called the dampening factor, is usually set between 0.1 and 0.2 [83]

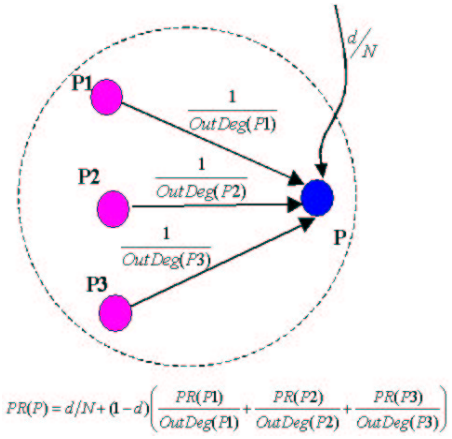

Fig. 2. PageRank -Markov Model for Random Web Surfer

a link. Figure 2 illustrates this concept , by showing how the PageRank of the page p is calculated.

- **Hubs and Authorities**: Hubs and Authorities can be viewed as 'fans' and 'centers' in a bipartite core of a Web graph. This is depicted in Figure 3, where the nodes on the left represent the hubs and the nodes on the right represent the authorities. The hub and authority scores computed for each Web page indicate the extent to which the Web page serves as a "hub" pointing to good "authority" pages or as an "authority" on a topic pointed to by good hubs. The hub and authority scores are computed for a set of pages related to a topic using an iterative procedure called HITS [52]. First a query is submitted to a search engine and a set of relevant documents is retrieved. This set, called the 'root set', is then expanded by including Web pages that point to those in the 'root set' and are pointed by those in the 'root set'. This new set is called the 'Base Set'. An adjacency matrix, A is formed such that if there exists at least one hyperlink from page i to page j, then $A_{i,j} = 1$, otherwise $A_{i,j} = 0$.HITS algorithm is then used to compute the "hub and "authority" scores for these set of pages.

There have been modifications and improvements to the basic *PageRank* and *Hubs and Authorities* approaches such as SALSA [59], Topic Sensitive PageRank [42] and Web page Reputations [65]. These different hyperlink based metrics have been discussed in [75].

3.2 Robot Detection and Filtering - Separating human and non-human Web behavior

Web robots are software programs that automatically traverse the hyperlink structure of the Web to locate and retrieve information. The importance of

Fig. 3. Hubs and Authorities

separating robot behavior from human behavior prior to building user behavior models has been illustrated by [80]. First of all, e-commerce retailers are particularly concerned about the unauthorized deployment of robots for gathering business intelligence at their Web sites. In addition, Web robots tend to consume considerable network bandwidth at the expense of other users. Sessions due to Web robots also make it difficult to perform click-stream analysis effectively on the Web data. Conventional techniques for detecting Web robots are often based on identifying the IP address and user agent of the Web clients. While these techniques are applicable to many well-known robots, they are not sufficient to detect camouflaged and previously unknown robots. [93] proposed an approach that uses the navigational patterns in click-stream data to determine if it is due to a robot. Experimental results have shown that highly accurate classification models can be built using this approach. Furthermore, these models are able to discover many camouflaged and previously unidentified robots.

3.3 Information scent - Applying foraging theory to browsing behavior

Information scent is a concept that uses the snippets and information presented around the links in a page as a "scent" to evaluate the quality of content of the page it points to, and the cost of accessing such a page [21]. The key idea is to model a user at a given page as "foraging" for information,and following a link with a stronger "scent". The "scent" of a path depends on how likely it is to lead the user to relevant information , and is determined by a network flow algorithm called spreading activation. The snippets, graphics, and other information around a link are called "proximal cues". The user's desired information need is expressed as a weighted keyword vector. The simi-

larity between the proximal cues and the user's information need is computed as "Proximal Scent". With the proximal cues from all the links and the user's information need vector, a "Proximal Scent Matrix" is generated. Each element in the matrix reflects the extent of similarity between the link's proximal cues and the user's information need. If enough information is not available around the link, a "Distal Scent" is computed with the information about the link described by the contents of the pages it points to. The "Proximal Scent" and the "Distal Scent" are then combined to give the "Scent" Matrix. The probability that a user would follow a link is decided by the "scent" or the value of the element in the "Scent" matrix. Figure 4 depicts a high level view of this model. Chi et al. [21] proposed two new algorithms called Web User Flow by Information Scent (WUFIS) and Inferring User Need by Information Scent (IUNIS) using the theory of information scent based on Information foraging concepts. WUFIS predicts user actions based on user needs, and IUNIS infers user needs based on user actions.The concept is illustrated in 4.

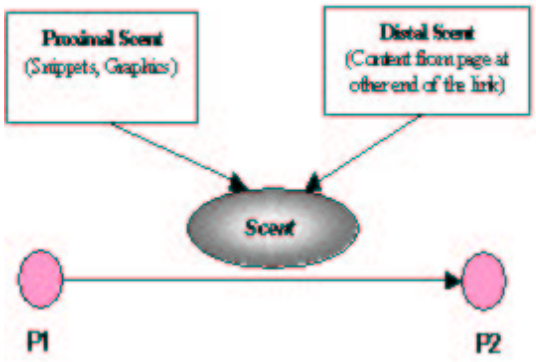

Fig. 4. Information Scent

3.4 User profiles - Understanding how users behave

The Web has taken user profiling to completely new levels. For example, in a 'brick-and-mortar' store, data collection happens only at the checkout counter, usually called the 'point-of-sale'. This provides information only about the final outcome of a complex human decision making process, with no direct information about the process itself. In an on-line store, the complete click-stream is recorded, which provides a detailed record of every single action taken by the user, providing a much more detailed insight into the decision making process. Adding such behavioral information to other kinds of information about users,

e.g. demographic, psychographic, etc., allows a comprehensive user profile to be built, which can be used for many different applications [64].

While most organizations build profiles of user behavior limited to visits to their own sites, there are successful examples of building 'Web-wide' behavioral profiles, e.g. Alexa Research [2] and DoubleClick [31]. These approaches require browser cookies of some sort, and can provide a fairly detailed view of a user's browsing behavior across the Web.

3.5 Interestingness measures - When multiple sources provide conflicting evidence

One of the significant impacts of publishing on the Web has been the close interaction now possible between authors and their readers. In the pre-Web era, a reader's level of interest in published material had to be inferred from indirect measures such as buying/borrowing, library checkout/renewal, opinion surveys, and in rare cases feedback on the content. For material published on the Web it is possible to track the precise click-stream of a reader to observe the exact path taken through on-line published material.We can measure exact times spent on each page, the specific link taken to arrive at a page and to leave it, etc. Much more accurate inferences about readers' interest in content can be drawn from these observations. Mining the user click-stream for user behavior, and using it to adapt the 'look-and-feel' of a site to a reader's needs was first proposed in [69].

While the usage data of any portion of a Web site can be analyzed, the most significant, and thus 'interesting', is the one where the usage pattern differs significantly from the link structure. This is interesting because the readers' behavior, reflected by Web usage, is very different from what the author would like it to be - reflected by the structure created by the author. Treating knowledge extracted from structure data and usage data as evidence from independent sources, and combining them in an evidential reasoning framework to develop measures for interestingness has been proposed in [9,28].

3.6 Pre-processing - making Web data suitable for mining

In the panel discussion referred to earlier [91] , pre-processing of Web data to make it suitable for mining was identified as one of the key issues for Web mining. A significant amount of work has been done in this area for Web usage data, including user identification [79] , session creation [79] , robot detection and filtering [93] , extracting usage path patterns [89], etc. Cooley's Ph.D. thesis [28] provides a comprehensive overview of the work in Web usage data preprocessing.

Preprocessing of Web structure data, especially link information, has been carried out for some applications, the most notable being Google style Web search [83]. An up-to-date survey of structure preprocessing is provided in [75].

3.7 Topic Distillation

Topic Distillation is the identification of a set of documents or parts of document that are most relevant to a given topic. It has been defined [15]as

> 'the process of finding authoritative Web pages and comprehensive 'hubs' which reciprocally endorse each other and are relevant to a given query.'

Kleinberg 's HITS approach [52] was one of early link based approach that addressed the issue of identifying Web pages related to a specific topic. Bharath and Henzinger [8] and Chakrabarti et al [13, 24] used hyperlink analysis to automatically identify the set of documents relevant to a given topic. Katz and Li [95] used a three step approach - (i) Document Keyword Extraction, (ii) Keyword propagation across pages connected by links, and (iii) keyword propagation through category tree structure - to automatically distill topics from the set of documents belonging to a category or to extract documents related to certain topics. The FOCUS project [17,18,37] concentrates on building portals pertaining to a topic automatically. A 'fine-grained model' based on the Document Object Model (DOM) of a page and the hyperlink structure of hubs and authorities related to a topic has also been developed [16]. This approach reduces topic drift and helps in identifying parts of a Web page relevant to a query.

In recent work on identifying topics, Mendelzon and Rafiei [65] define a new measure called 'reputation' of a page and compute the set of topics for which a page will be rated high. Haveliwala [42] proposed a 'Topic-Sensitive PageRank', which pre-computes a set of PageRank vectors corresponding to different topics.

3.8 Web Page Categorization

Web page categorization determines the category or class a Web page belongs to, from a pre-determined set of categories or classes. Topic Distillation is similar but in Web page categorization, the categories can be based on topics or other functionalities, e.g. home pages, content pages, research papers, etc, whereas Topic Distillation is concerned mainly with content-oriented topics. Pirolli et al [76] defined a set of 8 categories for nodes representing Web pages and identified 7 different features based on which a Web page could be categorized into these 8 categories. Chakrabarti et al [15] use a relaxation labeling technique to model a class-conditional probability distribution for assigning a category by looking at the neighboring documents that link to the given document or linked by the given the document. Attardi et al [5] proposed an automatic method of classifying Web pages based on the link and context. The idea is that if a page is pointed to by another page, the link would carry certain context weight since it induces someone to read the given page from the page that is referring to it. Getoor et al [60] treat documents

and links as entities in an Entity- Relationship model and use a Probabilistic Relational Model to specify the probability distribution over the document-link database, and classify the documents using belief propagation methods. Chakrabarti et al [14]describe how topic taxonomies and automatic classifiers can be used to estimate the distribution of broad topics in the whole Web.

3.9 Identifying Web Communities of information sources

The Web has had tremendous success in building communities of users and information sources. Identifying such communities is useful for many purposes.We discuss here a few significant efforts in this direction.

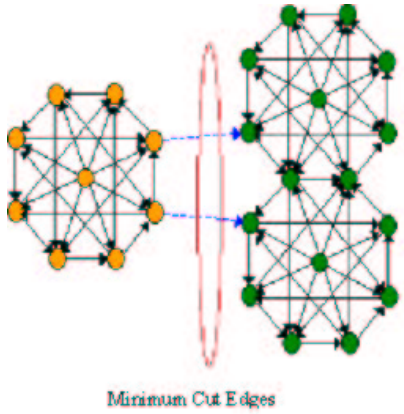

Fig. 5. Maximal Flow Model for Web Commmunities

Gibson et al. [35] identified Web communities as "*a core of central 'authoritative' pages linked together by 'hub' pages*".Their approach was extended by Ravi Kumar et al. in [54] to discover emerging Web communities while crawling. A different approach to this problem was taken by Flake et al [38] who applied the "maximum-flow minimum cut model" [48] to the Web graph for identifying "Web communities". This principle is illustrated in Figure 5. Imafuji et al. [70]compare the HITS and the maximum flow approaches and discuss the the strengths and weakness of the two methods. Reddy et al. [77] propose a dense bipartite graph method, a relaxation to the complete bipartite method followed by HITS approach, to find Web communities.A related concept of "Friends and Neighbors" was introduced by Adamic and Adar in [56]. They identified a group of individuals with similar interests, who in the cyber-world would form a "community". Two people are termed "friends" if the similarity between their Web pages is high. The similarity is measured using the features:*text, out-links, in-Links and mailing lists.*

3.10 Online Bibiliometrics

With the Web having become the fastest growing and most upto date source of information, the research community has found it extremely useful to have online repository of publications. Lawerence et al. have observed in [86] that having articles online makes them more easily accessible and hence more often cited than articles that were offline. Such online repositories not only keep the researchers updated on work carried out at different centers , but also makes the interaction and exchange of information much easier.The concept is illustrated in Figure 6

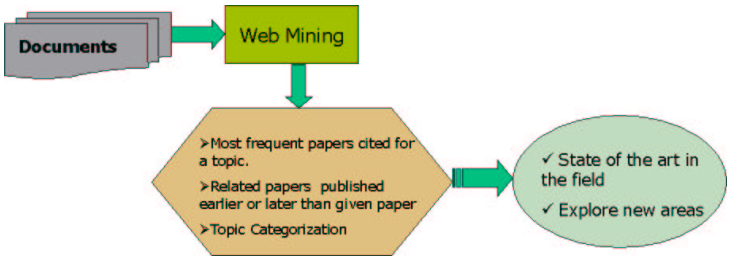

Fig. 6. Information Extraction in Online Bibiliometrics

With such information stored in the Web, it becomes easier to point to the most frequent papers that are cited for a topic and also related papers that have been published earlier or later than a given paper. This helps in understanding the 'state of the art' in a particular field, helping researchers to explore new areas. Fundamental Web mining techniques are applied to improve the search and categorization of research papers, and citing related articles. Some of the prominent digital libraries are SCI [85], ACM portal [1], CiteSeer [23] and DBLP [30].

3.11 Semantic Web Mining

The data in the World Wide Web is largely in an unstructured format which makes information retrieval and navigation on the Web a difficult task. Automatic retrieval of information is an even more daunting task as the large amount of data available in the web is written mostly for human interpretation with no semantic structure attached. With the amount of information overload on the internet, it is necessary to develop automatic agents that can perform the challenging task of extracting information from the web. Existing search engines apply their own heuristics to arrive at the most relevant web pages for a query. Though they are found to be very useful, there is lack of preciseness as the search engines are not able to identify the exact semantics of the documents. Hence, there is a need for a more structured semantic document that would help in better retrieval and exchange of information.

At the highest level Semantic Web can be thought as adding certain semantic structures to the existing Web data. Semantic Web is also closely related to other areas such as Semantic Networks [98] and Conceptual graphs [88]that have been extensively studied and have been adopted to the web domain. In Semantic Networks, the edges of such a graph represent the semantic relationship between the vertices. Among other techniques used for Semantic Web is the RDF and XML Topic Maps. RDF data consists of nodes and attached attribute/value pairs that can be modeled as labelled directed graphs. Topic maps are used to organise large amount of information in an optimal way for better management and navigation. Topic maps can be viewed as the online versions of printed indices and catalogs. Topic Maps are essentially network of the topics that can be formed using the semantics from the underlying data. Tim Berners Lee [58] from W3C describes best the idea behind having such structured well-defined documents as:

> 'The concept of machine-understandable documents does not imply some magical artificial intelligence which allows machines to comprehend human mumblings. It only indicates a machine's ability to solve a well-defined problem by performing well-defined operations on existing well-defined data. Instead of asking machines to understand people's language, it involves asking people to make the extra effort.'

Web Mining techniques can be applied to learn ontologies for the vast source of unstructured web data available. Doing this manually for the whole web is definitely not scalable or practical. Conversely, defining ontologies for existing and future documents will help in faster and more accurate retrieval of documents. Berendt et al [7] discuss in more detail about the integration of the two topics -'Semantic Web' and 'Web Mining'.

3.12 Visualization of the World Wide Web

Mining Web data provides a lot of information, which can be better understood with visualization tools. This makes concepts clearer than is possible with pure textual representation. Hence,there is a need to develop tools that provide a graphical interface that aids in visualizing results of Web mining.

Analyzing the web log data with visualization tools has evoked a lot of interest in the research community. Chi et al. in [20] developed a Web Ecology and Evolution Visualization (WEEV) tool to understand the relationship between Web content, Web structure and Web Usage over a period of time. The site hierarchy is represented in a circular form called the "Disk Tree" and the evolution of the Web is viewed as a 'Time Tube'.Cadez et al. in [12] present a tool called WebCANVAS that displays clusters of users with similar navigation behavior. Prasetyo et al. in [10] introduce N̈aviz;, a interactive web log visualization tool that is designed to display the user browsing pattern on the web site at a global level and then display each browsing path on the pattern displayed earlier in an incremental manner. The support of each traversal is

represented by the thickness of the edge between the pages. The user browsing path that is of interest can be displayed by specifying the pages, or the number of intermediate nodes that have been traversed to reach a page. Such a tool is very useful in analyzing user behavior and improving web sites.

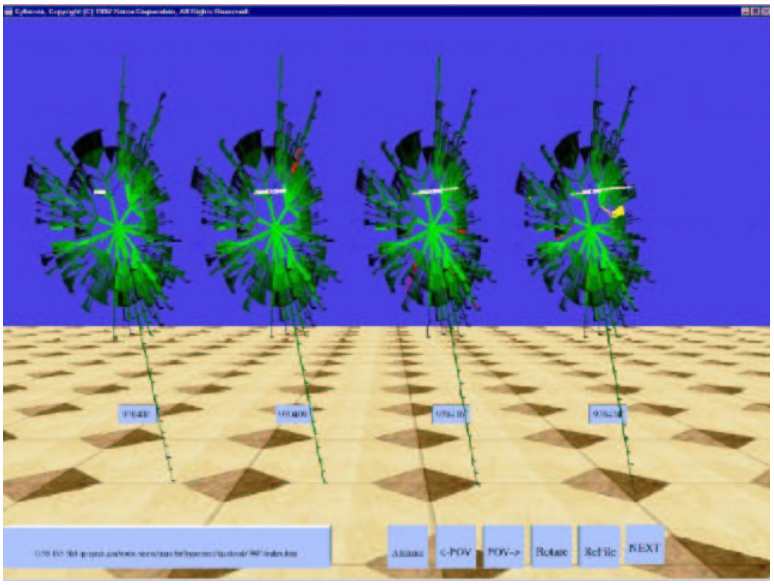

Fig. 7. Time Tube consisting of four disk trees representing evolution of Web Ecology. Figure taken from [20]

4 PROMINENT APPLICATIONS

An outcome of the excitement about the Web in the past few years has been that Web applications have been developed at a much faster rate in the industry than research in Web related technologies. Many of these are based on the use of Web mining concepts,even though the organizations that developed these applications, and invented the corresponding technologies, did not consider it as such. We describe some of the most successful applications in this section. Clearly, realizing that these applications use Web mining is largely a retrospective exercise.For each application category discussed below, we have selected a prominent representative, purely for exemplary purposes. This in no way implies that all the techniques described were developed by that organization alone. On the contrary, in most cases the successful techniques were developed by a rapid 'copy and improve' approach to each other's ideas.

4.1 Personalized Customer Experience in B2C E-commerce - Amazon.com

Early on in the life of Amazon.com, its visionary CEO Jeff Bezos observed,

> 'In a traditional (brick-and-mortar) store, the main effort is in getting a customer to the store. Once a customer is in the store they are likely to make a purchase - since the cost of going to another store is high - and thus the marketing budget (focused on getting the customer to the store) is in general much higher than the in-store customer experience budget (which keeps the customer in the store). In the case of an on-line store, getting in or out requires exactly one click, and thus the main focus must be on customer experience in the store.'[2]

This fundamental observation has been the driving force behind Amazon's comprehensive approach to personalized customer experience, based on the mantra 'a personalized store for every customer' [68]. A host of Web mining techniques, e.g. associations between pages visited, click-path analysis, etc., are used to improve the customer's experience during a 'store visit'. Knowledge gained from Web mining is the key intelligence behind Amazon's features such as 'instant recommendations', 'purchase circles', 'wish-lists', etc. [3].

Fig. 8. Amazon.com's personalized Web page

[2] The truth of this fundamental insight has been borne out by the phenomenon of 'shopping cart abandonment', which happens frequently in on-line stores, but practically never in a brick-and-mortar one.

4.2 Web Search - Google

Google [41] is one of the most popular and widely used search engines. It provides users access to information from over 2 billion web pages that it has indexed on its server. The quality and quickness of the search facility, makes it the most successful search engine. Earlier search engines concentrated on Web content alone to return the relevant pages to a query. Google was the first to introduce the importance of the link structure in mining the information from the web. PageRank, that measures the importance of a page, is the underlying technology in all Google search products,and uses structural information of the Web graph to return high quality results.

The 'Google Toolbar' is another service provided by Google that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages. The full version of the toolbar, if installed, also sends the click-stream information of the user to Google. The usage statistics thus obtained is used by Google to enhance the quality of its results. Google also provides advanced search capabilities to search images and find pages that have been updated within a specific date range. Built on top of Netscape's Open Directory project, Google's web directory provides a fast and easy way to search within a certain topic or related topics. The Advertising Programs introduced by Google targets users by providing advertisements that are relevant to a search query. This does not bother users with irrelevant ads and has increased the clicks for the advertising companies by four or five times. According to BtoB, a leading national marketing publication, Google was named a top 10 advertising property in the Media Power 50 that recognizes the most powerful and targeted business-to-business advertising outlets [39].

One of the latest services offered by Google is,' Google News' [40]. It integrates news from the online versions of all newspapers and organizes them categorically to make it easier for users to read "the most relevant news". It seeks to provide latest information by constantly retrieving pages from news site worldwide that are being updated on a regular basis. The key feature of this news page, like any other Google service, is that it integrates information from various Web news sources through purely algorithmic means, and thus does not introduce any human bias or effort. However, the publishing industry is not very convinced about a fully automated approach to news distillations [90].

4.3 Web-wide tracking - DoubleClick

'Web-wide tracking', i.e. tracking an individual across all sites he visits is one of the most intriguing and controversial technologies. It can provide an understanding of an individual's lifestyle and habits to a level that is unprecedented, which is clearly of tremendous interest to marketers. A successful example of this is DoubleClick Inc.'s DART ad management technology [31]. DoubleClick

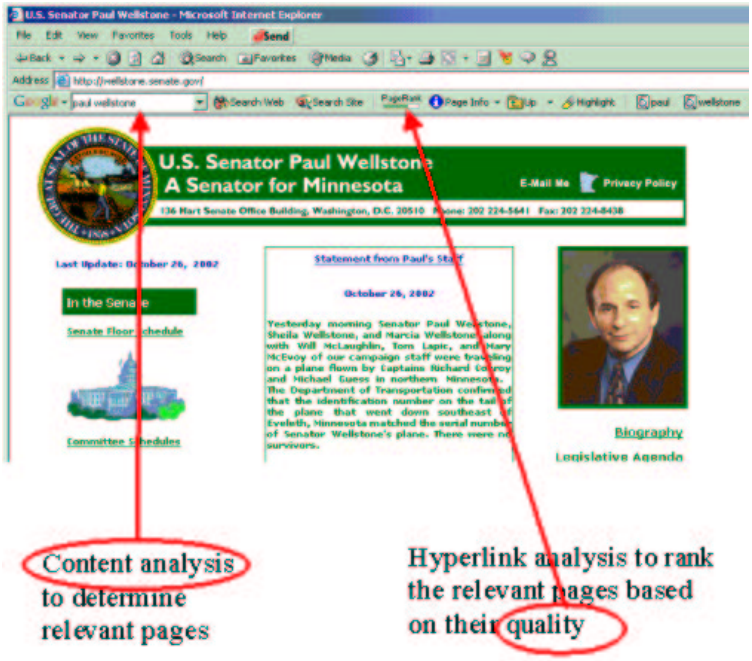

Fig. 9. Web page returned by Google for query "Paul Wellstone"

serves advertisements, which can be targeted on demographic or behavioral attributes, to end-user on behalf of the client, i.e. the Web site using DoubliClick's service. Sites that use DoubleClick's service are part of 'The DoubleClick Network' and the browsing behavior of a user can be tracked across all sites in the network, using a cookie. This makes DoubleClick's ad targeting to be based on very sophisticated criteria. Alexa Research [2] has recruited a panel of more than 500,000 users, who have voluntarily agreed to have their every click tracked, in return for some freebies. This is achieved through having a browser bar that can be downloaded by the panelist from Alexa's website, which gets attached to the browser and sends Alexa a complete click-stream of the panelist's Web usage. Alexa was purchased by Amazon for its tracking technology.

Clearly Web-wide tracking is a very powerful idea. However, the invasion of privacy it causes has not gone unnoticed, and both Alexa/Amazon and DoubleClick have faced very visible lawsuits [32, 34]. Microsoft's "Passport" technology also falls into this category [66]. The value of this technology in applications such a cyber-threat analysis and homeland defense is quite clear, and it might be only a matter of time before these organizations are asked to provide this information to law enforcement agencies.

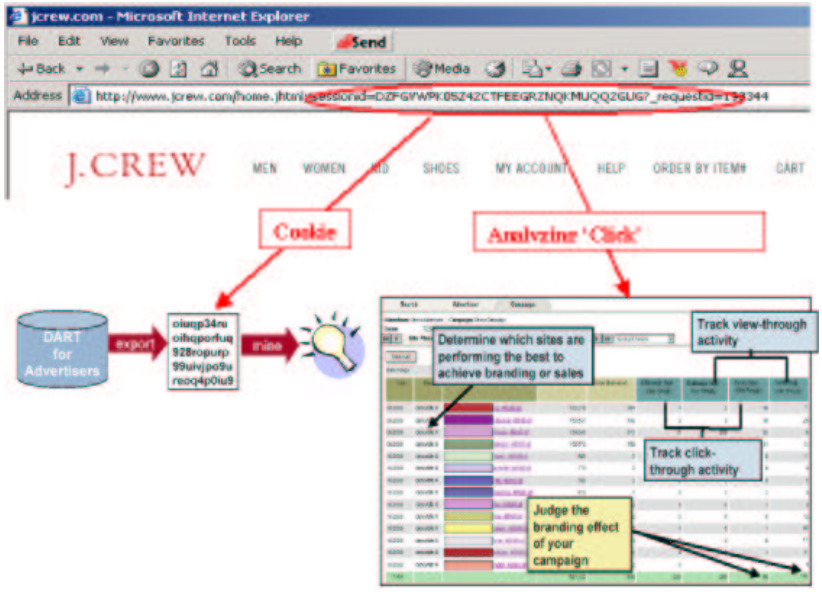

Fig. 10. DART system for Advertisers, DoubleClick

4.4 Understanding Web communities - AOL

One of the biggest successes of America Online (AOL) has been its sizeable and loyal customer base [4]. A large portion of this customer base participates in various 'AOL communities', which are collections of users with similar interests. In addition to providing a forum for each such community to interact amongst themselves, AOL provides them with useful information and services. Over time these communities have grown to be well-visited 'waterholes' for AOL users with shared interests. Applying Web mining to the data collected from community interactions provides AOL with a very good understanding of its communities, which it has used for targeted marketing through ads and e-mail solicitation. Recently, it has started the concept of 'community sponsorship', whereby an organization, say Nike, may sponsor a community called 'Young Athletic TwentySomethings'. In return, consumer survey and new product development experts of the sponsoring organization get to participate in the community, perhaps without the knowledge of other participants. The idea is to treat the community as a highly specialized focus group, understand its needs and opinions on new and existing products; and also test strategies for influencing opinions.

4.5 Understanding auction behavior - eBay

As individuals in a society where we have many more things than we need, the allure of exchanging our 'useless stuff' for some cash, no matter how

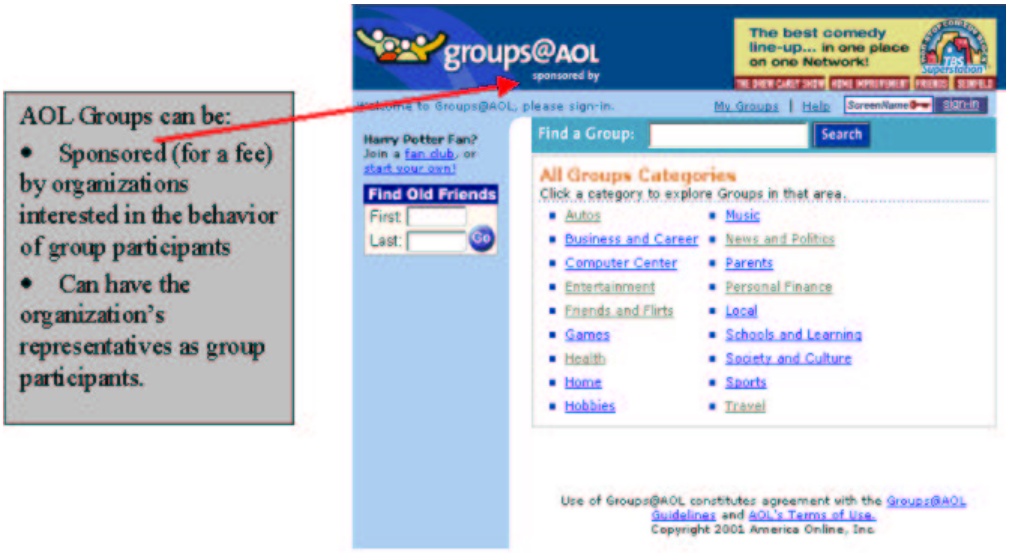

Fig. 11. Groups at AOL: Understanding user community

small, is quite powerful. This is evident from the success of flea markets, garage sales and estate sales. The genius of eBay's founders was to create an infrastructure that gave this urge a global reach, with the convenience of doing it from one's home PC [36]. In addition, it popularized auctions as a product selling/buying mechanism, which provides the thrill of gambling without the trouble of having to go to Las Vegas. All of this has made eBay as one of the most successful businesses of the Internet era. Unfortunately, the anonymity of the Web has also created a significant problem for eBay auctions, as it is impossible to distinguish real bids from fake ones. eBay is now using Web mining techniques to analyze bidding behavior to determine if a bid is fraudulent [25]. Recent efforts are towards understanding participants' bidding behaviors/patterns to create a more efficient auction market.

4.6 Personalized Portal for the Web - MyYahoo

Yahoo [99] was the first to introduce the concept of a 'personalized portal', i.e. a Web site designed to have the look-and-feel and content personalized to the needs of an individual end-user. This has been an extremely popular concept and has led to the creation of other personalized portals, e.g. Yodlee [100] for private information, e.g bank and brokerage accounts.

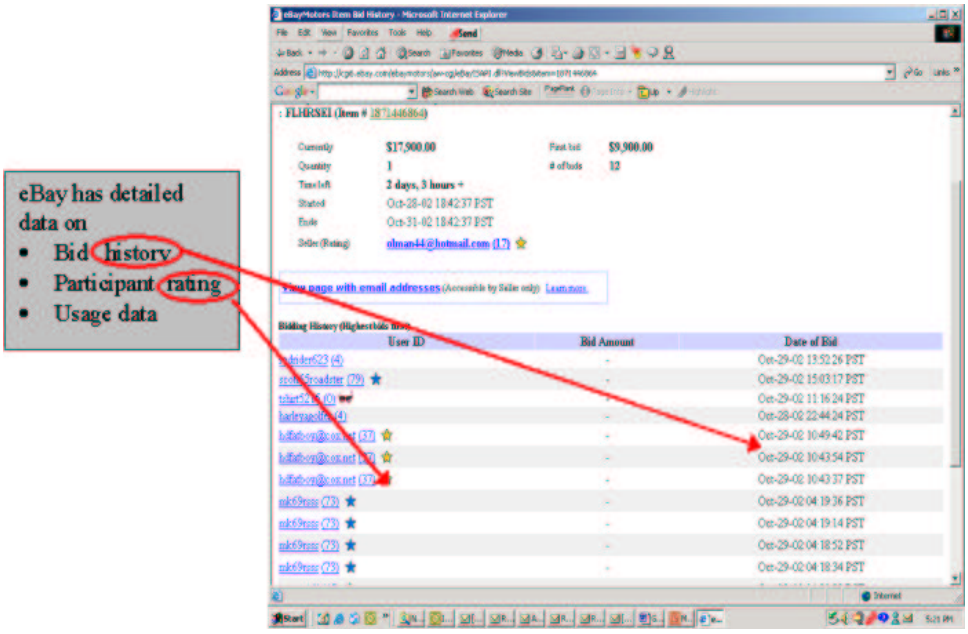

Fig. 12. E-Bay: Understanding Auction Behavior

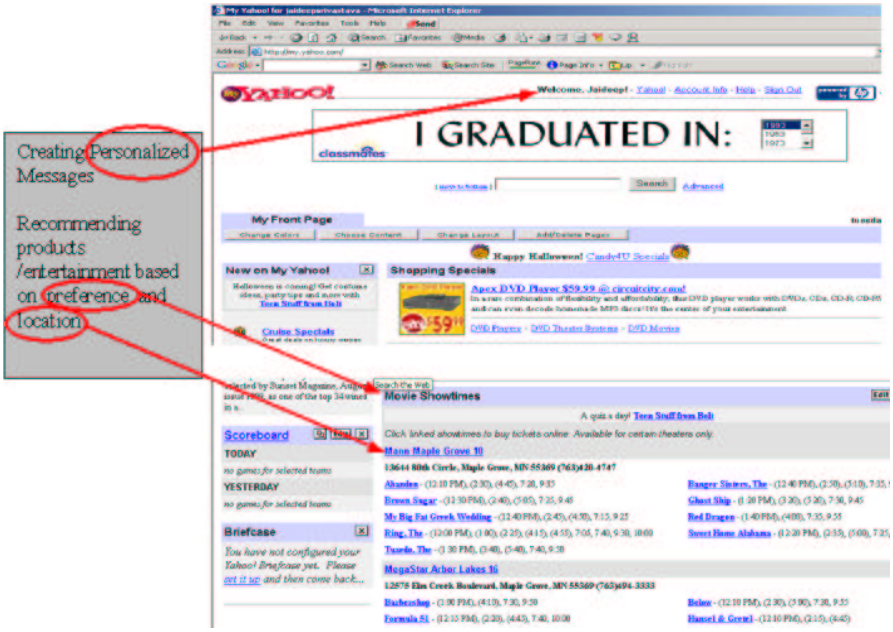

Fig. 13. My Yahoo: Personalized Webpage

Mining MyYahoo usage logs provides Yahoo valuable insight into an individual's Web usage habits, enabling Yahoo to provide personalized content, which in turn has led to the tremendous popularity of the Yahoo Web site.[3]

4.7 CiteSeer - Digital Library and Autonomous Citation Indexing

NEC ResearchIndex,also known as CiteSeer [23, 51], is one of the most popular online bibiliographic indices related to Computer Science. The key contribution of the CiteSeer repository is the "Autonomous Citation Indexing" (ACI) [87]. Citation indexingmakes it possible to extract information about related articles. Automating such a process reduces a lot of human effort,and makes it more effective and faster. The key concepts are depicted in Figure 14.

Fig. 14. CiteSeer - Autonomous Citation Indexing

CiteSeer works by crawling the Web and downloading research related papers. Information about citations and the related context is stored for each of these documents.The entire text and information about the document is stored in different formats. Information about documents that are similar at a sentence level (percentage of sentences that match between the documents), at a text level or related due to co-citation is also given. Citation statistics for documents are computed that enable the user to look at the most cited or popular documents in the related field. They also a maintain a directory for

[3] Yahoo has been consistently ranked as one of the top Web property for a number of years [67].

computer science related papers , to make search based on categories easier. These documents are ordered by the number of citations.

4.8 i-MODE: Accessing the Web through Cell phones

i-MODE is a cell-phone service from NTT DoCoMo, Japan [43]. It has about 40 million users who access the internet from their cell phones. The internet connections are continuous and the customers can access the specially tailored web sites as long as they are in the area that receives the i-mode the signal. Unlike the 'circuit-switched' based systems that require dial-up, i-mode is 'packet-switched' and hence continuous. This enables the download of information from the web sites to the cell phones faster, and without having to worry about the connection time. Users can receive and send email, do online shopping or banking, stock trading, receive traffic news and weather forecasts, and search for restaurants and other local places.

Fig. 15. I-MODE: NTT DoCoMo's mobile internet acccess system. The figure is taken from the The Eurotechnology Website [43]

The usual speed for i-mode download ranges from 28.8 kbit/sec for top range models to the order of 200 kbit/sec for FOMA (3rd Generation) services. As a markup language, I-mode uses cHTML (compact HTML), which is in an extended subset of ordinary HTML that concentrates on text and simple graphics. The i-mode markup language also has certain special i-mode only tags and image characters that are used as emoticon symbols. The size of an i-mode page is limited to 5 kbytes. These set of web pages open a new domain of information. they have their own structure, semantics and usage. The content of these pages are alo restricted depending on the needs of the end users. Such a domain provides usage data based on an individual and this can be very useful to identify interesting user behavior patterns.

4.9 OWL: Web Ontology Language

The OWL Web Ontology Language [72] is designed for automatic processing of web content information to derive meaningful contexts without any human intervention. OWL has three expressive sublanguages: OWL Lite, OWL DL, and OWL Full. OWL is used to express the ontological information - meanings and relationships among different words - from the Web content.

One succesful application of such a language is a web portal. The web portal has an a single web page that is used as a starting point to search among the listed topics on the page. The list of topics and the topics associated with a web page is usally done manually and submitted. However, there has also been extensive research on automatic 'topic distillation' and 'web page categorization'. OntoWeb and Open Directory Projects are typical examples of such portals.

4.10 vTag Web Mining Server- Connotate Technologies

Connotate Technogies [26] was founded by Data Mining and Machine Learning Scientists at Rutger's Univeristy. They are the developers of Web Services products that help users to browse and convert information from unstructured documents on the Web to a more structured format like XML. This conversion helps in providing better insight for personalizations, business intelligence and other enterprise solutions. The overall architecture of the the vTag Web Mining Server can be seen in Figure 16

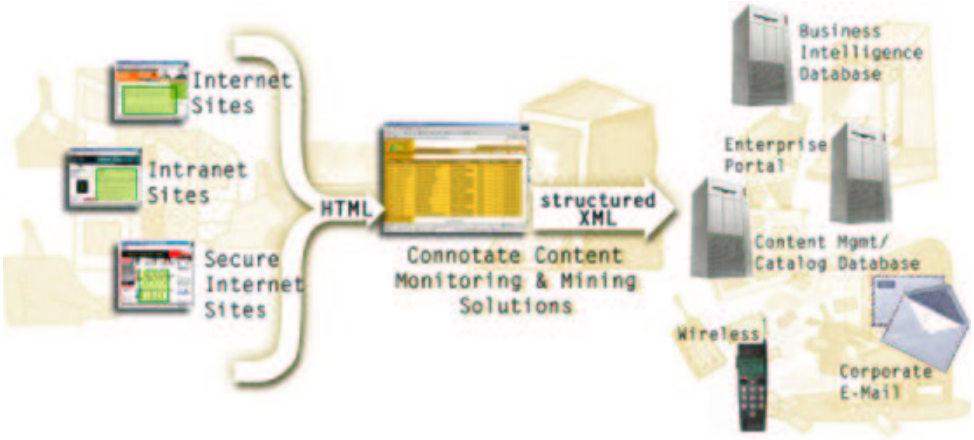

Fig. 16. vTag Web MiningServer Architecture. Figure taken from white paper on Web content and services mining [27]

The Web Mining Server supports information agents that monitor, extract and summarise the information from the various web sources. These informa-

tion agents are easy and quick to set up using a a graphical user interface. The user can set it up according to the fatures they they think are essential to keep track of. There is no special need for programmers. The automation of this process helps busineeses and enterprises to better track the neccesary information from the large amount of web pages and summarise them for further analysis and action. Information agents are also capable of converting unstructured data iinto structured form and store it in a database. Creation of such information agents requires no special skills and can be done easily using the graphical user interface provided. The content that is converted to a structured format like XML can be used for business intelligence, supply chain integrations etc. The converted content can also be sent as an e-mail or a message to a user in his mobile.

5 RESEARCH DIRECTIONS

Even though we are going through an inevitable phase of 'irrational despair' following a phase of 'irrational exuberance' about the commercial potential of the Web, the adoption and usage of the Web continues to grow unabated [96]. As the Web and its usage grows, it will continue to generate evermore content, structure, and usage data, and the value of Web mining will keep increasing. Outlined here are some research directions that must be pursued to ensure that we continue to develop Web mining technologies that will enable this value to be realized.

5.1 Web metrics and measurements

From an experimental human behaviorist's viewpoint, the Web is the perfect experimental apparatus. Not only does it provides the ability of measuring human behavior at a micro level, it eliminates the bias of the subjects knowing that they are participating in an experiment, and allows the number of participants to be many orders of magnitude larger than conventional studies. However, we have not yet begun to appreciate the true impact of a revolutionary experimental apparatus for human behavior studies. The Web Lab of Amazon [3] is one of the early efforts in this direction. It is regularly used to measure the user impact of various proposed changes - on operational metrics such as site visits and visit/buy ratios, as well as on financial metrics such as revenue and profit - before a deployment decision is made. For example, during Spring 2000 a 48 hour long experiment on the live site was carried out, involving over one million user sessions, before the decision to change Amazon's logo was made. Research needs to be done in developing the right set of Web metrics, and their measurement procedures, so that various Web phenomena can be studied.

5.2 Process mining

Mining of 'market basket' data, collected at the point-of-sale in any store, has been one of the visible successes of data mining. However, this data provides only the end result of the process, and that too decisions that ended up in product purchase. Click-stream data provides the opportunity for a detailed look at the decision making process itself, and knowledge extracted from it can be used for optimizing the process, influencing the process, etc. [97]. Underhill [78] has conclusively proven the value of process information in understanding users' behavior in traditional shops. Research needs to be carried out in (i) extracting process models from usage data, (ii) understanding how different parts of the process model impact various Web metrics of interest, and (iii) how the process models change in response to various changes that are made, i.e. changing stimuli to the user. Figure 17 shows an approach of modeling online shopping as a state transition diagram.

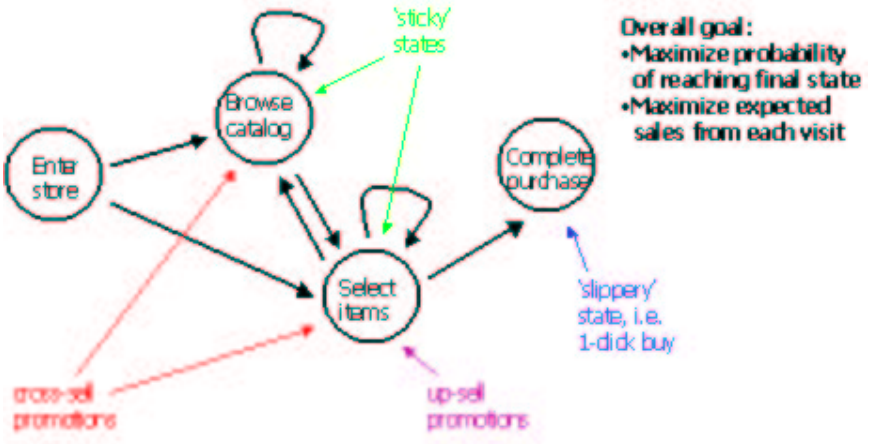

Fig. 17. Shopping Pipeline modeled as State Transition Diagram

5.3 Temporal evolution of the Web

Society's interaction with the Web is changing the Web as well as the way people interact. While storing the history of all of this interaction in one place is clearly too staggering a task, at least the changes to the Web are being recorded by the pioneering Internet Archive project [44]. Research needs to be carried out in extracting temporal models of how Web content, Web structures, Web communities, authorities, hubs, etc. evolve over time. Large organizations generally archive (at least portions of) usage data from there Web sites. With these sources of data available, there is a large scope of

research to develop techniques for analyzing of how the Web evolves over time. Figure 18 shows how content, structure and usage of Web information can evolve over time.

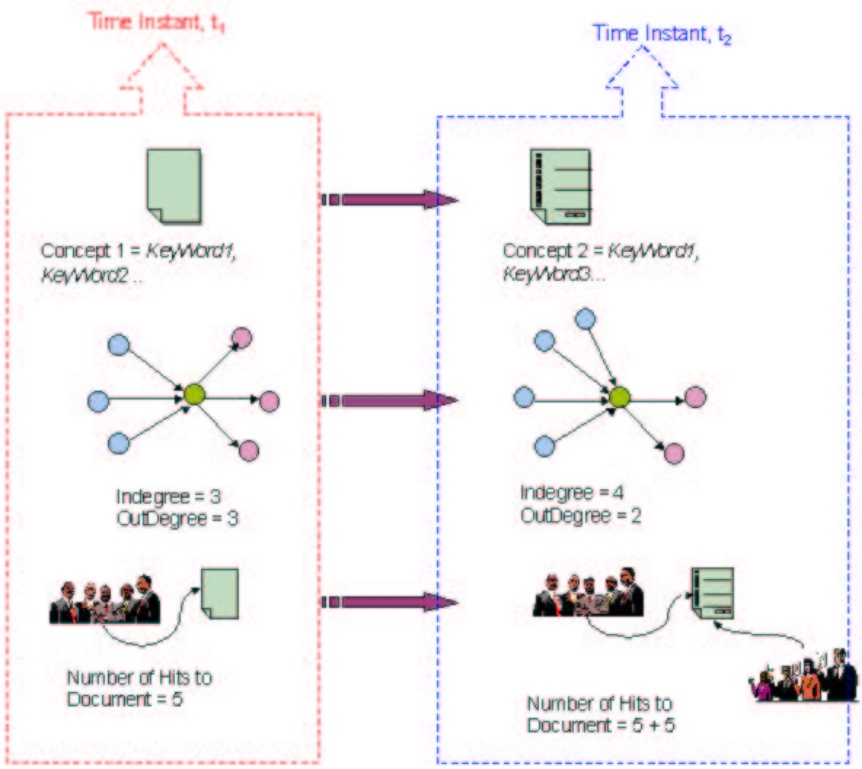

Fig. 18. Temporal Evolution for a single document in the World Wide Web

5.4 Web services optimization

As services over the Web continue to grow [50], there will be a continuing need to make them robust, scalable and efficient. Web mining can be applied to better understand the behavior of these services, and the knowledge extracted can be useful for various kinds of optimizations. The successful application of Web mining for predictive pre-fetching of pages by a browser has been demonstrated in [73]. Research is needed in developing Web mining techniques to improve various other aspects of Web services.

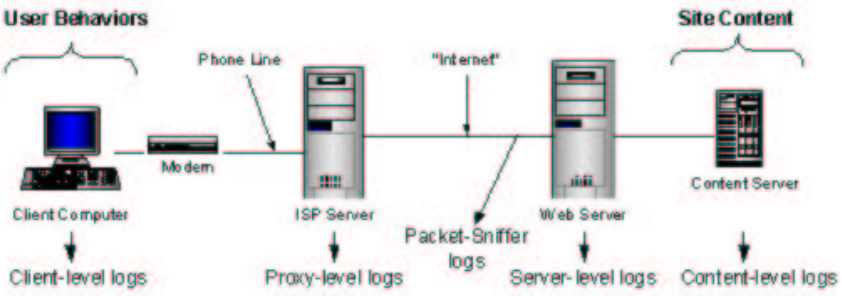

Fig. 19. High Level Architecture of Different Web Services

5.5 Distributed Web Mining

The data on the World Wide Web is huge and distributed across different web sites. To analuyse such a data one can integrate all the data to one site and perform the required analysis. However, such an approach is time consuming and not scalable. A better approach would be to analyse the data locally at the different locations and build an overall model. This can be done in two diffrent ways: surreptious and co-operative. Surreptious approaches are those in which the user behavior across different web sites is tracked and integrated without the user explicitly having to submit any information. In the more co-operative approaches like the D-DOS attacks, such unusual behavior is centrally reported to the CERT organisation. Chen et al [19] have developed bayesian network model for mining web user logs from multiple sites.

Also with the increasing use of wireless netorks and accessing of of the internet through the cell phones, a large amount of usage information can be collected across different web server logs. With such information interesting user behavioral patterns can be mined. Personalization of the Web sites depending on the user locations and interests would be more effective analysing such data. Thus a concept of 'Life on the Web' for an individual can defined by integrating such information. Hence there is a need to develop models for the distributed data and efficient integration for extracting useful information. The data distributed across different servers could have different nature and also the computing resources at different locations may vary.Hence, there is a need to develop different web mining algorithms to extract useful models [74].

5.6 Mining Information from E-mails-Discovering evolving user trends

E-mails have found to contain a huge amount information both in terms of its content, its usage and the evolving network built by sending e-mails. Target marketing using e-mails is one field that has been proved to be very effective according to a recent survey done by DoubleClick [33]. E-mails are a big

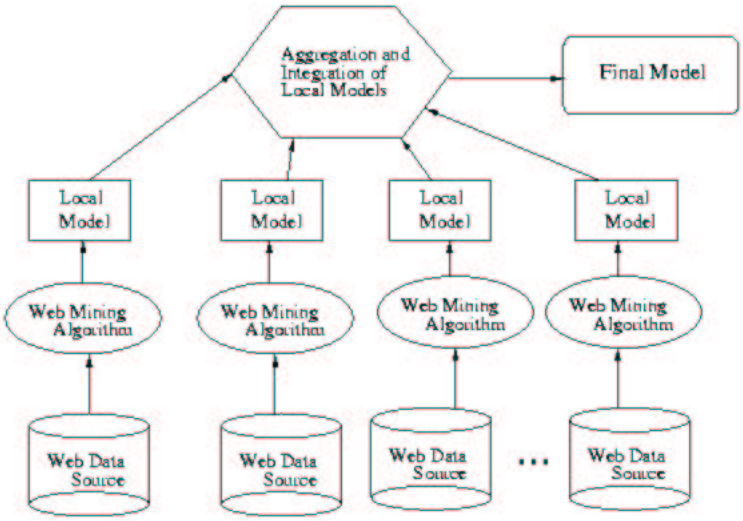

Fig. 20. Distributed Web Mining Framework. Adopted from [74]

source for multi-channel purchases. E-mails provide very useful information to track user interests and purchasing behavior and helps in increasing the level of personalization that can be offered. For example, according to the survey conducted by DoubleClick, women are more receptive to promotions and discounts and their idea of spam differes from that of men. And hence e-mail can serve as an excellent online source for marketing products related to women. However, e-mail faces the problem of spam that annoys users. Both consumers and companies providing free e-mail services are using tools to limit the spam to make the web life of the user more comfortable. Limiting spam helps in removing the removing the 'noise' from the data provided by e-mails.

While marketing is one such area where E-mail provides an excellent source of information, this could be extended to other areas too. For example, there are E-mail groups that consists of prospective graduate students. Such groups would provide an excellent feedback of what the student interests are and what are the universities that are popular. Mining this kind of information would be useful for Universities that give admissions and also for porspective

students in the later years. Mining information from such e-mail groups would help to uderstand the user needs and the underlying trends.

E-mails also form a network of directed graphs with the nodes being people or a work position(like operator@, webmaster@) and an e-mail sent from one one node to the other represented as a directed edge. This kind of a network is also dynamic and possesses a temporal dimension. The basic structure os the netwrok and its usage will be very useful in providing information about online communties, significant persons or abnormal user behavior.

5.7 Fraud and threat analysis

The anonymity provided by the Web has led to a significant increase in attempted fraud, from unauthorized use of individual credit cards to hacking into credit card databases for blackmail purposes [84]. Yet another example is auction fraud, which has been increasing on popular sites like eBay [USDoJ2002]. Since all these frauds are being perpetrated through the Internet, Web mining is the perfect analysis technique for detecting and preventing them. Research issues include developing techniques to recognize known frauds, and characterize and then recognize unknown or novel frauds, etc. The issues in cyber threat analysis and intrusion detection are quite similar in nature [57].

5.8 Web mining and privacy

While there are many benefits to be gained from Web mining, a clear drawback is the potential for severe violations of privacy. Public attitude towards privacy seems to be almost schizophrenic, i.e. people say one thing and do quite the opposite. For example, famous case like [34] and [31] seem to indicate that people value their privacy, while experience at major e-commerce portals shows that over 97% of all people accept cookies with no problems - and most of them actually like the personalization features that can be provided based on it. Spiekerman et al [92] have demonstrated that people were willing to provide fairly personal information about themselves, which was completely irrelevant to the task at hand, if provided the right stimulus to do so. Furthermore, explicitly bringing attention to information privacy policies had practically no effect. One explanation of this seemingly contradictory attitude towards privacy may be that we have a bi-modal view of privacy, namely that "I'd be willing to share information about myself as long as I get some (tangible or intangible) benefits from it, as long as there is an implicit guarantee that the information will not be abused". The research issue generated by this attitude is the need to develop approaches, methodologies and tools that can be used to verify and validate that a Web service is indeed using an end-user's information in a manner consistent with its stated policies.

6 CONCLUSIONS

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper we have briefly described the key computer science contributions made by the field, a number of prominent applications, and outlined some promising areas of future research. Our hope is that this overview provides a starting point for fruitful discussion.

7 ACKNOWLEDGEMENTS

The ideas presented here have emerged in discussions with a number of people over the past few years - far too numerous to list. However, special mention must be made of Robert Cooley, Mukund Deshpande, Joydeep Ghosh, Ronny Kohavi, Ee-Peng Lim, Brij Masand, Bamshad Mobasher, Ajay Pandey, Myra Spiliopoulou, Pang-Ning Tan, Terry Woodfield, and Masaru Kitsuregawa discussions with all of whom have helped develop the ideas presented herein. This work was supported in part by the Army High Performance Computing Research Center contract number DAAD19-01-2-0014. The ideas and opinions expressed herein do not necessarily reflect the position or policy of the government (either stated or implied) and no official endorsement should be inferred. The AHPCRC and the Minnesota Super-computing Institute provided access to computing facilities.

References

1. ACM Portal. `http://portal.acm.org/portal.cfm`.
2. Alexa research. `http://www.alexa.com`.
3. Amazon.com. `http://www.amazon.com`.
4. America Online. `http://www.aol.com`, 2002.
5. Giuseppe Attardi, Antonio Gullí, and Fabrizio Sebastiani. Automatic Web page categorization by link and context analysis. In Chris Hutchison and Gaetano Lanzarone, editors, *Proceedings of THAI-99, European Symposium on Telematics, Hypermedia and Artificial Intelligence*, pages 105–119, Varese, IT, 1999.
6. BEA Weblogic Server. `http://www.bea.com/products/weblogic/server/index.shtml`.
7. B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining, 2002.
8. Krishna Bharat and Monika R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 104–111, Melbourne, AU, 1998.

9. B.Padmanabhan and A.Tuzhilin. A Belief-Driven Method for Discovering Unexpected Patterns. In *Knowledge Discovery and Data Mining*, pages 94–100, 1998.
10. B.Prasetyo, I. Pramudiono, K. Takahashi, M.Toyoda, and M.Kitsuregawa. Naviz user behavior visualization of dynamic page.
11. Broadvision 1-to-1 portal. `http://www.bvportal.com/`.
12. I.V. Cadez, D.Heckerman, C.Meek, P.Smyth, and S.White. Visualization of navigation patterns on a Web site using modelbased clustering. In *Knowledge Discovery and Data Mining*, pages 280–284, 2000.
13. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*, 1998.
14. S. Chakrabarti, M. Joshi, K. Punera, and D. Pennock. The structure of broad topics on the web, 2002.
15. Soumen Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *World Wide Web*, pages 211–220, 2001.
16. Soumen Chakrabarti, Mukul Joshi, and Vivek Tawde. Enhanced topic distillation using text, markup tags, and hyperlinks. In *Research and Development in Information Retrieval*, pages 208–216, 2001.
17. Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Distributed hypertext resource discovery through examples. In *The VLDB Journal*, pages 375–386, 1999.
18. Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1623–1640, 1999.
19. R. Chen, K. Sivakumar, and H. Kargupta. Distributed web mining using Bayesian networks from multiple data streams. In *Proceedings of ICDM 2001*, pages 75–82, 2001.
20. E.H. Chi, J.Pitkow, J.Mackinlay, P.Pirolli, R.Gossweiler, and S.K. Card. Visualizing the evolution of web ecologies. In *Proceedings of the Conference on Human Factors in Computing Systems CHI'98*, 1998.
21. E.H. Chi, P.Pirolli, K.Chen, and J.E. Pitkow. Using Information Scent to model user information needs and actions and the Web. In *Proceedings of CHI 2001*, pages 490–497, 2001.
22. C.H.Moh, E.P.Lim, and W.K.Ng. DTD-Miner: A Tool for Mining DTD from XML Documents. WECWIS, 2000.
23. CiteSeer Scientific Literature Digital Library. `http://citeseer.nj.nec.com/cs`.
24. The CLEVER Project. `http://www.almaden.com/cs/k53/clever.html`.
25. E. Colet. Using Data Mining to Detect Fraud in Auctions, 2002.
26. Connotate technologies. `http://www.connotate.com/`.
27. Web Services Content Mining: Extract, Monitor and Deliver. `http://www.connotate.com/web_mining_white_paper.pdf`.
28. R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data, 2000.
29. R. Cooley, J. Srivastava, and B. Mobasher. Web Mining: Information and Pattern Discovery on the World Wide Web, 1997.

30. DBLP Bibiliography. http://www.informatik.uni-trier.de/~ley/db/.
31. DoubleClick's DART Technology. http://www.doubleclick.com/dartinfo/, 2002.
32. DoubleClick's Lawsuit. http://www.wired.com/news/business/0,1367,36434,00.html, 2002.
33. DoubleClick 2003 Consumer Email Study . http://www.doubleclick.com/us/knowledge_central/documents/research/dc_c%onsumeremailstudy_0310.pdf, 2003.
34. C. Dembeck and P. A. Greenberg. Amazon: Caught Between a Rock and a Hard Place. http://www.ecommercetimes.com/perl/story/2467.html, 2002.
35. D.Gibson, J.M. Kleinberg, and P. Raghavan. Inferring Web Communities from Link Topology. In *UK Conference on Hypertext*, pages 225–234, 1998.
36. eBay Inc. http://www.ebay.com.
37. The FOCUS project. http://www.cs.berkeley.edu/~soumen/focus/.
38. G.Flake, S.Lawrence, and C.L.Giles. Efficient Identification of Web Communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 20–23 2000.
39. Google Recognized As Top Business-To-Business Media Property. http://www.google.com/press/pressrel/b2b.html.
40. Google News. http://news.google.com.
41. Google Inc. http://www.google.com.
42. T. Haveliwala. Topic-sensitive PageRank. In Proceedings of the Eleventh International World Wide Web Conference, Honolulu, Hawaii, May 2002., 2002.
43. I-mode. http://http://www.eurotechnology.com/imode/index.html.
44. The Internet Archive Project. http://www.archive.org/.
45. J.Borges and M.Levene. Mining Association Rules in Hypertext Databases. In *Knowledge Discovery and Data Mining*, pages 149–153, 1998.
46. J.Ghosh and J. Srivastava. Proceedings of Workshop on Web Analytics. http://www.lans.ece.utexas.edu/workshop_index2.htm, 2001.
47. J.Ghosh and J. Srivastava. Proceedings of Workshop on Web Mining. http://www.lans.ece.utexas.edu/workshop_index.htm, 2001.
48. L.R. Ford Jr and D.R. Fulkerson. Maximal Flow through a network, 1956.
49. 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):12–23, 2000.
50. R.H. Katz. Pervasive Computing: It's All About Network Services, 2002.
51. K.Bollacker, S.Lawrence, and C.L.Giles. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In Katia P. Sycara and Michael Wooldridge, editors, *Proceedings of the Second International Conference on Autonomous Agents*, pages 116–123, New York, 1998. ACM Press.
52. J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
53. R. Kohavi, M. Spiliopoulou, and J. Srivastava. Proceedings of WebKDD2000 - Web Mining for E-Commerce - Challenges and Opportunities, 2001.
54. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1481–1493, 1999.

55. K.Wang and H.Liu. Discovering Typical Structures of Documents: A Road Map Approach. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–154, 1998.
56. L.Adamic and E.Adar. Friends and Neighbors on the Web.
57. A. Lazarevic, P. Dokas, L. Ertoz, V. Kumar, J. Srivastava, and P.N. Tan. Data mining for network intrusion detection, 2002.
58. Tim Berners Lee. What the semantic web can represent. `http://www.w3.org/DesignIssues/RDFnot.html`, 1998.
59. R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks (Amsterdam, Netherlands: 1999)*, 33(1–6):387–401, 2000.
60. L.Getoor, E.Segal, B.Tasker, and D.Koller. Probabilistic models of text and link structure for hypertext classification. IJCAI Workshop on Text Learning:Beyond Supervision, Seattle, WA, 2001.
61. L.Page, S.Brin, R.Motwani, and T.Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
62. S.K. Madria, S.S. Bhowmick, W.K Ng, and E.P Lim. Research Issues in Web Data Mining. In *Data Warehousing and Knowledge Discovery*, pages 303–312, 1999.
63. B. Masand and M. Spiliopoulou. Proceedings of WebKDD1999 - Workshop on Web Usage Analysis and User Profiling, 1999.
64. B. Masand, M. Spiliopoulou, J Srivastava, and O. Zaiane. Proceedings of WebKDD2002 - Workshop on Web Usage Patterns and User Profiling, 2002.
65. A.O. Mendelzon and D. Rafiei. What do the Neighbours Think? Computing Web Page Reputations. *IEEE Data Engineering Bulletin*, 23(3):9–16, 2000.
66. MicroSoft.NET Passport. `http://www.microsoft.com/netservices/passport/`.
67. Top 50 US Web and Digital Properties. `http://www.jmm.com/xp/jmm/press/mediaMetrixTop50.xml`.
68. E. Morphy. Amazon pushes 'personalized store for every customer'. `http://www.ecommercetimes.com/perl/story/13821.html`, 2001.
69. M.Perkowitz and O.Etzioni. Adaptive Web Sites: Conceptual Cluster Mining. In *IJCAI*, pages 264–269, 1999.
70. N.Imafuji and M.Kitsuregawa. Effects of maximum flow algorithm on identifying web community. In *Proceedings of the fourth international workshop on Web information and data management*, pages 43–48. ACM Press, 2002.
71. O.Etzioni. The World-Wide Web: Quagmire or Gold Mine? *Communications of the ACM*, 39(11):65–68, 1996.
72. OWL. `http://www.w3.org/TR/owl-features/`.
73. A. Pandey, J.Srivastava, and S.Shekhar. A web intelligent prefetcher for dynamic pages using association rules - a summary of results, 2001.
74. Byung-Hoon Park and Hillol Kargupta. Distributed data mining: Algorithms, systems, and applications.
75. P.Desikan, J.Srivastava, V.Kumar, and P.N. Tan. Hyperlink Analysis-Techniques & Applications. Technical Report 2002-152, Army High Performance Computing Research Center, 2002.
76. Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: Extracting usable structures from the web. In *Proc. ACM Conf. Human Factors in Computing Systems, CHI*. ACM Press, 1996.

77. P.K.Reddy and M. Kitsuregawa. An approach to build a cyber-community hierarchy. Workshop on Web Analytics,held in Conjunction with Second SIAM International Conference on Data Mining, 2002.
78. P.Underhill. Why we buy: The Science of shopping. Touchstone Books, 2000.
79. R.Cooley, B.Mobasher, and J.Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999.
80. R.Kohavi. Mining e-commerce data: The good, the bad, and the ugly. In Foster Provost and Ramakrishnan Srikant, editors, *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 8–13, 2001.
81. R.Kohavi, B.Masand, M.Spiliopoulou, and J.Srivastava. Proceedings of WebKDD2001 - Mining Log Data Across All Customer Touchpoints, 2001.
82. R.Kosala and H.Blockeel. Web mining research: A survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery& Data Mining, ACM*, 2, 2000.
83. S.Brin and L.Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.
84. D. Scarponi. Blackmailer Reveals Stolen Internet Credit Card Data. `http://abcnews.go.com/sections/world/DailyNews/internet000110.html`, 2000.
85. Science Citation Index. `http://www.isinet.com/isi/products/citation/sci/`.
86. S.Lawrence. Online or invisible? *Nature*, 411(6837):521, 2001.
87. S.Lawrence, C.L.Giles, and K.Bollacker. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.
88. J. F. Sowa. Conceptual Structures Information Processing in Mind and Machine. . Addison Wesley, reading et al, 1984.
89. M. Spiliopoulou. Data Mining for the Web. Proceedings of the Symposium on Principles of Knowledge Discovery in Databases (PKDD), 1999.
90. T. Springer. Google LaunchesNews Service. `http://www.computerworld.com/developmenttopics/websitemgmt/story/0,1080%1,74470,00.html`, 2002.
91. J. Srivastava and B. Mobasher. Web Mining: Hype or Reality? . 9th IEEE International Conference on Tools With Artificial Intelligence (ICTAI '97), 1997.
92. S.Spiekermann, J.Grossklags, and B.Berendt. Privacy in 2nd generation E-Commerce: privacy preferences versus actual behavior. In *ACM Conference on Electronic Commerce*, pages 14–17, 2001.
93. P. Tan and V. Kumar. Discovery of web robot sessions based on their navigational patterns, 2002.
94. Vignette StoryServer. `http://www.cio.com/sponsors/110199\_vignette\_story2.html`.
95. V.Katz and W.S.Li. Topic Distillation in hierarchically categorized Web Documents. In *Proceedings of 1999 Workshop on Knowledge and Data Engineering Exchange, IEEE*, 1999.
96. Hosting Firm Reports Continued Growth. `http://thewhir.com/marketwatch/ser053102.cfm`, 2002.
97. K.L. Ong W.Keong. Mining Relationship Graphs for Eective Business Objectives.
98. W. Woods. What's in a link: Foundations for semantic networks. Academic Press, New York, 1975.

99. Yahoo!, Inc. http://www.yahoo.com.
100. Yodlee, Inc. http://www.yodlee.com.