Chapter 3

Web Mining - Accomplishments & Future Directions

Jaideep Srivastava[*], *Prasanna Desikan*[×], *Vipin Kumar*[†]
Department of Computer Science
200 Union Street SE, 4-192, EE/CSC Building
University of Minnesota, Minneapolis, MN 55455, USA
{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 existence, 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 **Keywords**: Web Mining,

3.1 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 [SM1997] asked the question "Is there anything distinct about Web mining (compared to data mining in general)?" While no

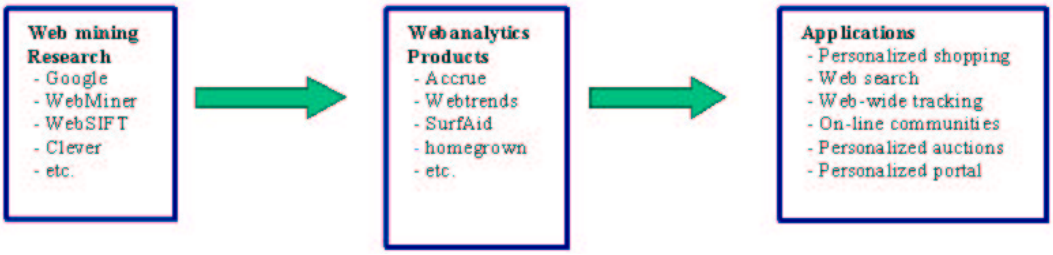


Figure 3.1: Web mining research & applications

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 [MS1999, KSS2000, KMSS2001, MSSZ2002], and the Web Analytics workshops, which have been held in conjunction with the SIAM data mining conference [GS2001, GS2002]. A good survey of the research in the field till the end of 1999 is provided in [KB2000] and [MBNL1999].

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 [E1996]. 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 [CMS1997]. The second definition has become more acceptable, as is evident from the approach adopted in most recent papers [MBNL1999, BL1999, KB2000] that have addressed the issue. In this paper we follow the data-centric view, and refine the definition of Web mining as,

> **Web mining** is the application of data mining techniques to extract knowledge from Web data, where **at least one of structure (hyperlink) or usage (Web log) data is used in the mining process** (with or without other types of Web data).

There is a purpose to adding the extra clause about structure and usage data. The reason being that mining Web content by itself is no different than general data mining, since it makes no difference whether the content was obtained from the Web, a database, a file system or through any other means. As shown in Figure 2, Web content can be variegated, containing text and hypertext, image, audio, video, records, etc. Mining each of these media types is by itself a sub-field of data mining.

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. One way to think about work in Web mining is as shown in Figure 3.1.

The rest of this paper is organized as follows : In section 3.2 we provide a taxonomy of Web mining, in section 3.3 we summarize some of the key results in the field, and in

section 4 we describe successful applications of Web mining techniques. In section 5 we present some directions for future research, and in section 6 we conclude the paper.

3.2 WEB MINING TAXONOMY

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:

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. Text mining and its application to Web content has been the most widely researched. Some of the research issues addressed in text mining are, topic discovery, extracting association patternss, clustering of web documents and classification of Web Pages. Research activities in this field also involve using techniques from 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 not been very rapid.

2. **Web Structure Mining:** The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be further divided into two kinds based on the kind of structural data used.

   - *Hyperlinks*: A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to 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 [DSKT2002] 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 [WL1998, MLN2000].

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 [SCDT2000]. 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:

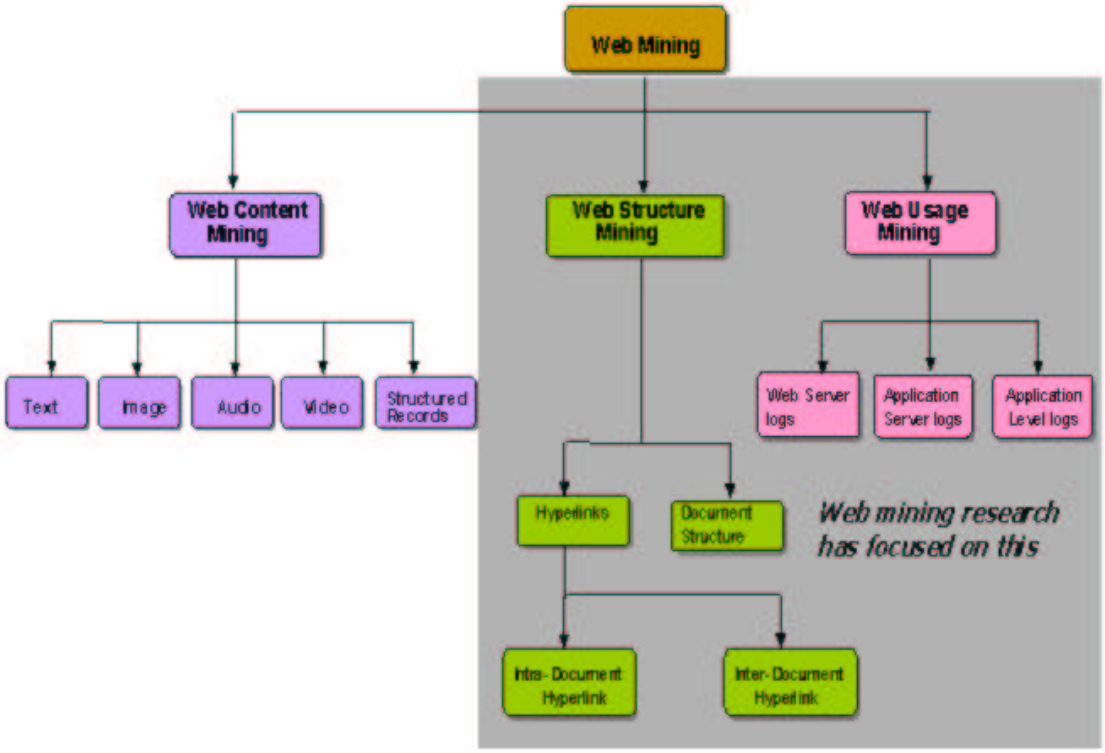


Figure 3.2: Web mining Taxonomy

- Web Server Data: They correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users.
- Application Server Data: Commercial application servers, e.g. Weblogic [BEA], BroadVision [BV], StoryServer [VIGN], 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 [SCDT2000].

3.3 KEY ACCOMPLISHMENTS

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

3.3.1 Page ranking metrics - Google's PageRank function

PageRank is a metric for ranking hypertext documents that determines the quality of these documents. Page et al. [PBMW1998] developed this metric for the popular search engine, Google [GOOGa, BP1998]. The key idea is that a page has 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 out of nowhere, i.e. (s)he could arrive at the page by typing the URL or from a bookmark, or may have a particular page as his/her homepage. d would then be the probability that a random surfer chooses a URL directly - i.e. typing it, using the bookmark list, or by default - rather than traversing a link[1] . Finally, 1/n corresponds to the uniform probability that a person chooses the page p from the complete set of n pages on the Web. The second term in the right hand side of the equation corresponds to factor contributed by arriving at a page by traversing a link. 1- d is the probability that a person arrives at the page p by traversing a link. The summation corresponds to the sum of the rank contributions made by all the pages that point to the page p. The rank contribution is the PageRank of the page multiplied by the probability that a particular link on the page is traversed. So for any page q pointing to page p, the probability that the link pointing to page p is traversed would be 1/OutDegree(q), assuming all links on the page is chosen with uniform probability. Figure 3.3(a) illustrates this concept clearly , by showing how the PageRank of the page P is calculated.

3.3.2 Hubs and Authorities - Identifying significant pages in the Web

Hubs and Authorities can be viewed as 'fans' and 'centers' in a bipartite core of a Web graph. A Core (i, j) is a complete directed bipartite sub-graph with at least i nodes from F and at least j nodes from C. With reference to the Web graph, i pages that contain the links are referred to as 'fans' and the j pages that are referenced are the 'centers'. From a conceptual point of view 'fans' and 'centers' in a Bipartite Core are basically

[1]The parameter d, called the dampening factor, is usually set between 0.1 and 0.2 [BP1998]

the Hubs and Authorities. This can be seen as depicted in Figure 3.3(b), 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 for a page are not based on a single formula, but are computed for a set of pages related to a topic using an iterative procedure called HITS algorithm [K1998]. We briefly give an overview of the procedure to obtain these scores. 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 whole 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 Ai, j = 1, otherwise Ai, j = 0. This computation is carried iteratively till the set does not expand further - or a threshold on iterations is reached.

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

Web robots are software programs that automatically traverse the hyperlink structure of the World Wide Web in order to locate and retrieve information. The importance of separating robot behavior from human behavior prior to extracting user behavior knowledge from usage data has been illustrated by [K2001]. 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 more 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 may not be sufficient to detect camouflaging and previously unknown robots. [TK2002] proposed an alternative approach that uses the navigational patterns in the 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 [TK2002]. Furthermore, these models are able to discover many camouflaging and previously unidentified robots.

3.3.4 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 to access such a page [CPCP2001]. The key idea is a user at a given page ”foraging” for information would follow a link with a stronger ”scent”. The ”scent” of the pages will decrease along a path and is determined by network flow algorithm called spreading activation. The snippets, graphics, and other information around a link are referred as ”proximal cues”. The user's desired information is expressed as

a weighted keyword vector. The similarity 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 3.3(c) depicts a high level view of this model. Chi et al [CPCP2001] 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 tends to predict user actions based on user needs and IUNIS infers user needs based on user actions.

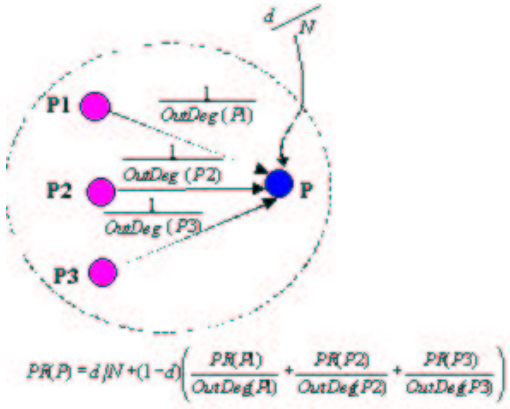
3.3.5 User profiles - Understanding how users behave

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 - which can provide 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 [MSSZ2002].

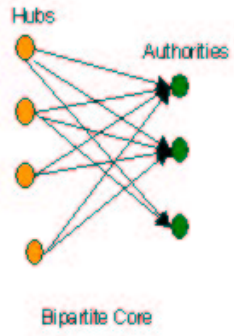
While most organizations build profiles of users' behavior limited to visits to their own sites, there are successful examples of building 'Web-wide' behavioral profiles, e.g. Alexa Research [ALEX] and DoubleClick [DCLKa]. These approaches require browser cookies of some sort, and can provide a fairly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

3.3.6 Interestingness measures

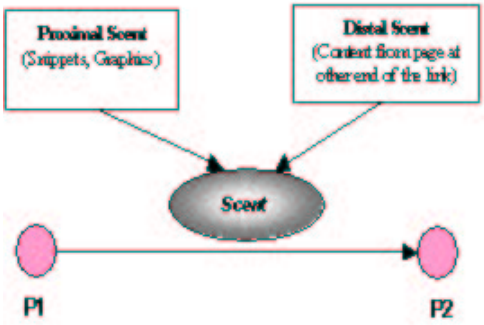
When multiple sources provide conflicting evidence One of the significant impacts of publishing on the Web has been the close interaction now possible between authors and their readers. In the pre-Web era, a reader's level of interest in published material had to be inferred from indirect measures such as buying/borrowing, library check-out/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, with 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 about published content can be drawn from these observations. Mining the user click-stream for user behavior, and use it to adapt the 'look-and-feel' of a site to a reader's needs was first proposed in [PE1999].



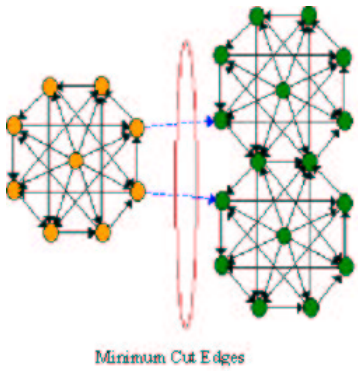
(a) PageRank



(b) Hubs and Authorities



(c) Information Scent



(d) Maximal Flow Model for Web Communities

Figure 3.3: Key Models developed in Web Mining

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 the 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 [PT1998, C2000].

3.3.7 Pre-processing - making Web data suitable for mining

In the panel discussion referred to above [SM1997], 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 [CMS1999, D1999], session creation [CMS1999, MSB2001], robot detection and filtering [TK2002], extracting usage path patterns [S1999], etc. Robert Cooley's Ph.D. thesis [C2000] provides a comprehensive overview of the work in Web usage data preprocessing.

Preprocessing of Web structure data, especially link data, has been carried out for some applications, the most notable being Google style Web search [BP1998]. An up-to-date survey of structure preprocessing is provided in [DSKT2002].

3.3.8 Maximum-Flow models - Web community identification

The idea of a maximal flow models has been used to identify communities, which can be described as a collection of Web pages such that each member node has more hyperlinks (in either direction) within the community than outside of the community. The s - t maximal flow problem can be described thus: Given a graph G = (V, E) whose edges are assigned positive flow capacities, and with a pair of distinguished nodes s and t, the problem is to find the maximum flow that can be routed from s to t. s is known as the source node and t as the sink node. Of course, the flow must strictly adhere to the constraints that arise due to the edge capacities. Ford and Fulkerson [FF1956] proposed that the maximal flow is equivalent to a "minimal cut" - that is the minimum number of edges that need to be cut from the graph to separate the source s from sink t. This principle is illustrated in Figure 3.3(d) Flake et al [FLG2000] have used this approach to identify "Web communities".

3.4 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 were 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.[2]

3.4.1 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.'[3]

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' [M2001]. 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. [AMZNa].

3.4.2 Web Search - Google

Google [GOOGa] is one of the most popular and widely used search engines. It provides users access to information from almost 2.5 billion web pages that it has indexed on its server. The simplicity and the quickness of the search facility, makes it the most successful search engine. Earlier search engines concentrated on the Web content 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 an importance of a page, is the underlying technology in all Google search products. The PageRank technology, that makes use of the structural information of the Web graph, is the key to returning quality results relevant to a query.

Google has successfully used the data available from the Web content (the actual text and the hyper-text) and the Web graph to enhance its search capabilities and provide best results to the users. Google has expanded its search technology to provide site-specific search to enable users to search for information within a specific website. The 'Google Toolbar' is another service provided by Google that seeks to make search

[2] 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.

[3] 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.

Figure 3.4: Amazon.com's personalized Web page

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 would be used by Google to enhance the quality of its results. Google also provides advanced search capabilities to search images and look for 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 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 [GOOGb].

One of the latest services offered by Google is,' Google News' [GOOGc]. 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 information that is the latest by constantly retrieving pages 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 [S2002].

Figure 3.5: Web page returned by Google for query “paul Wellstone”

3.4.3 Web-wide tracking - DoubleClick

’Web-wide tracking’, i.e. tracking an individual across all sites (s)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 - clearly of tremendous interest to marketers. A successful example of this is DoubleClick Inc.’s DART ad management technology [DCLKa]. DoubleClick serves advertisements, which can be targeted on demographic or behavioral attributes, to the 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 provides DoubleClick’s ad targeting to be based on very sophisticated criteria. Alexa Research [**?**] has recruited a panel of more than 500,000 users, who’ve 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 [DG2000, DCLKb]. 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.

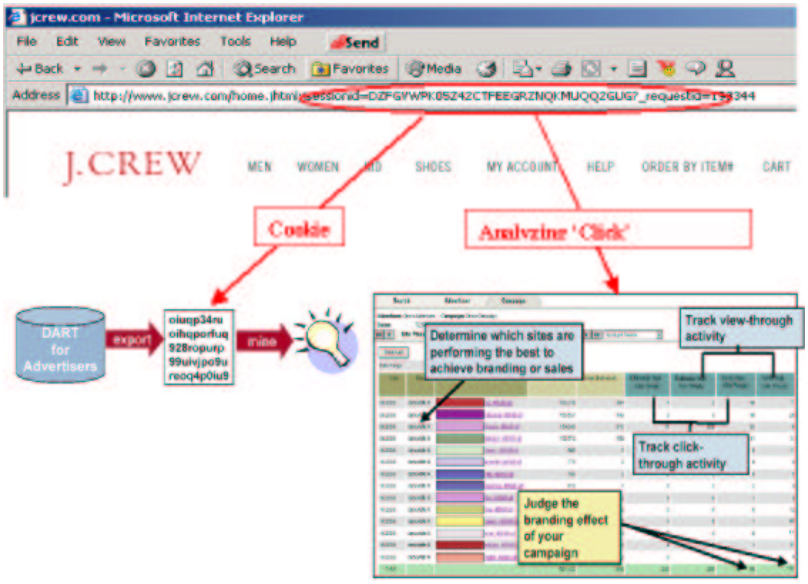


Figure 3.6: DART system for Advertisers, DoubleClick

3.4.4 Understanding Web communities - AOL

One of the biggest successes of America Online (AOL) has been its sizeable and loyal customer base [AOLa]. 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 useful information, etc. as well. 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 solicitations. Recently, it has started the concept of 'community sponsorship', whereby an organization like 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 - usually without the knowledge of the 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.

3.4.5 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 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

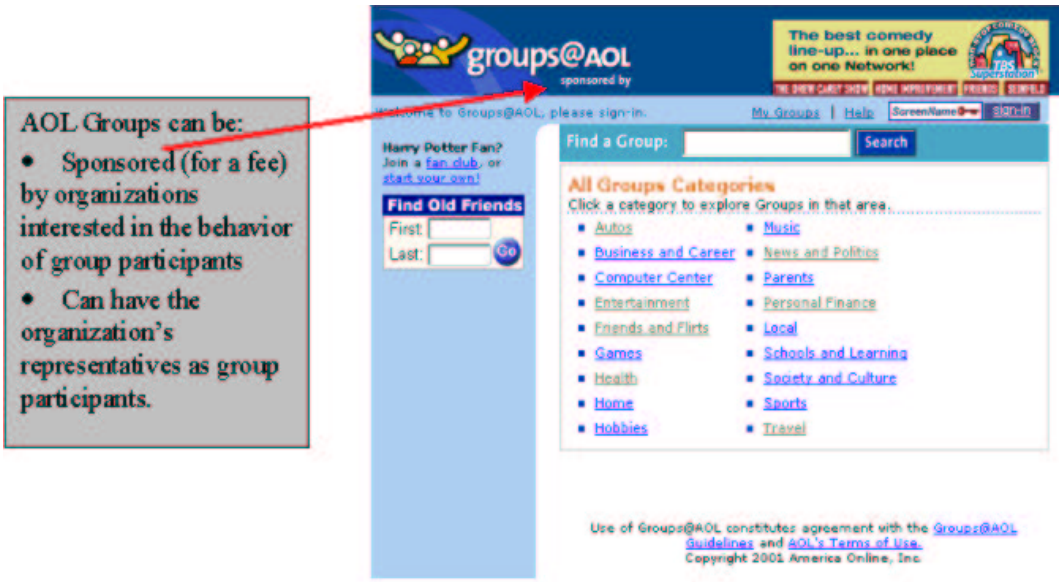


Figure 3.7: Groups at AOL: Understanding user community

reach, with the convenience of doing it from one's home PC [EBAYa]. 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 [C2002]. Recent efforts are towards understanding participants' bidding behaviors/patterns to create a more efficient auction market.

3.4.6 Personalized Portal for the Web - MyYahoo

Yahoo [YHOOa] was the first to introduce the concept of a 'personalized portal', i.e. a Web site designed to have the look-and-feel as well as 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 [YODLa] for private information. Mining MyYahoo usage logs provides Yahoo valuable insight into an individual's Web usage habits, enabling Yahoo to provide compelling personalized content, which in turn has led to the tremendous popularity of the Yahoo Web site.[4]

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

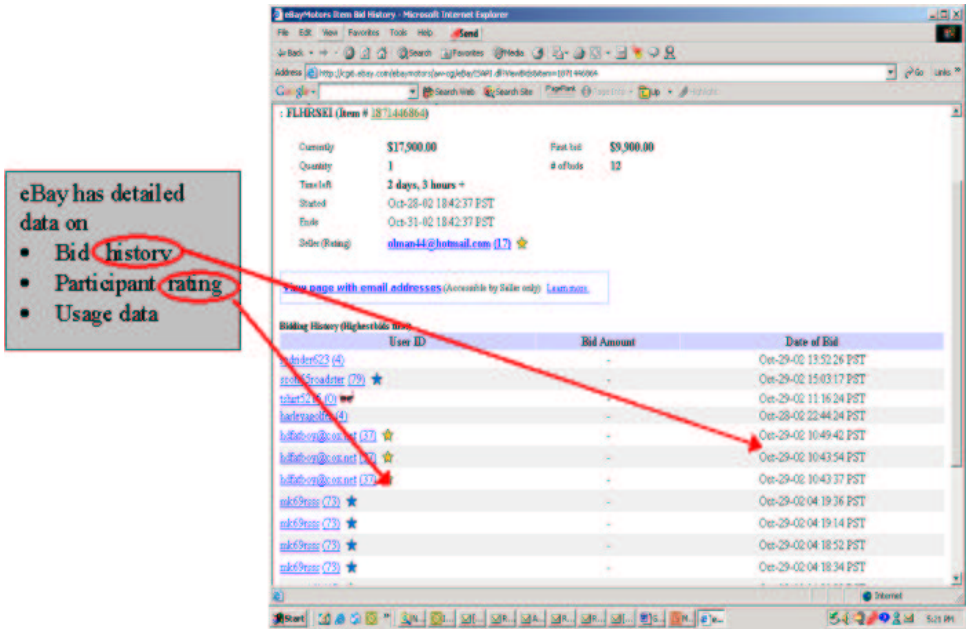


Figure 3.8: E-Bay: Understanding Auction Behavior

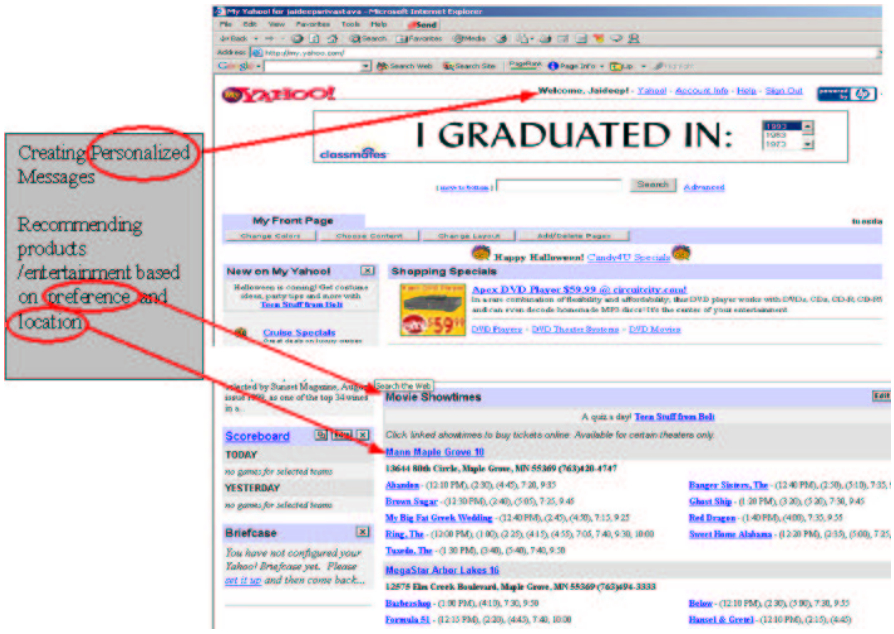


Figure 3.9: My Yahoo: Personalized Webpage

3.5 FUTURE DIRECTIONS

As we go 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 [WHN2002]. This trend is likely to continue as Web services continue to flourish [K2002]. 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.

3.5.1 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 (i) eliminates the bias of the subjects knowing that they are participating in an experiment, and (ii) allows the number of participants to be many orders of magnitude larger. However, we have not even begun to appreciate the true impact of a revolutionary experimental apparatus. The WebLab of Amazon [AMZNa] 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.

3.5.2 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. [ONL2002]. Underhill [U2000] 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 - changing stimuli to the user.

3.5.3 Temporal evolution of the Web

Society's interaction with the Web is changing the Web as well as the way the society interacts. While storing the history of all of this interaction in one place is clearly too

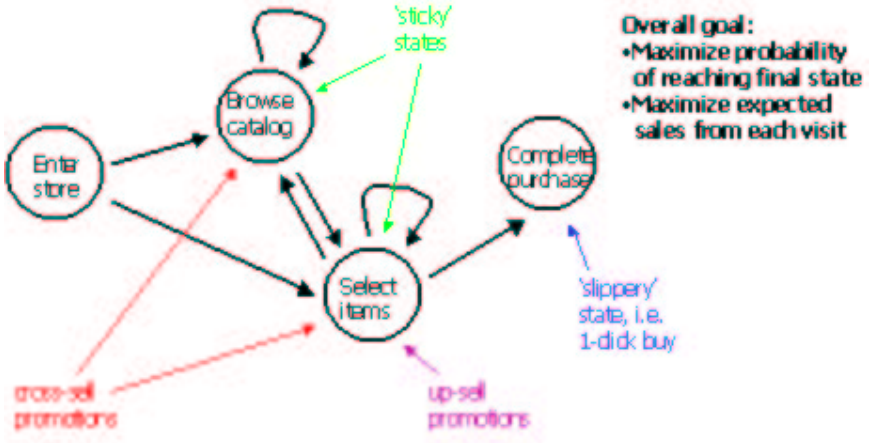


Figure 3.10: Shopping Pipeline modeled as State Transition Diagram

staggering a task, at least the changes to the Web are being recorded by the pioneering Internet Archive project [IA]. Research needs to be carried out in extracting temporal models of how Web content, Web structures, Web communities, authorities, hubs, etc. are evolving. 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.

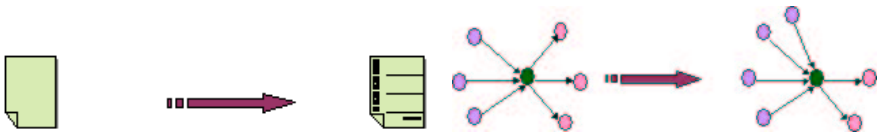
The temporal behavior of the three kinds of Web data: Web Content, Web Structure and Web Usage. The methodology suggested for Hyperlink Analysis in [DSKT2002] can be extended here and the research can be classified based on Knowledge Models, Metrics, Analysis Scope and Algorithms. For example, the analysis scope of the temporal behavior could be restricted to the behavior of a single document, multiple documents or the whole Web graph. The other factor that has to be studied is the effect of Web Content, Web Structure and Web Usage on each other over time.

3.5.4 Web services optimization

As services over the Web continue to grow [K2002], there will be a need to make them robust, scalable, efficient, etc. 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 [PSS2001]. Research is needed in developing Web mining techniques to improve various other aspects of Web services.

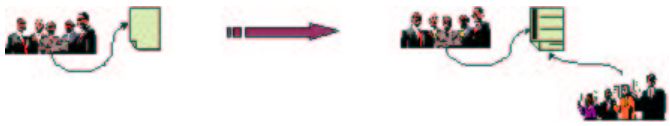
3.5.5 Fraud and threat analysis

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 [S2000]. Yet another example is auction fraud, which has been



(a) Change in Web Content of a document over time

(b) Change in Web Structure of a document over time



(c) Change in Web Usage of a document over time

Figure 3.11: Temporal Evolution for a single document in the World Wide Web

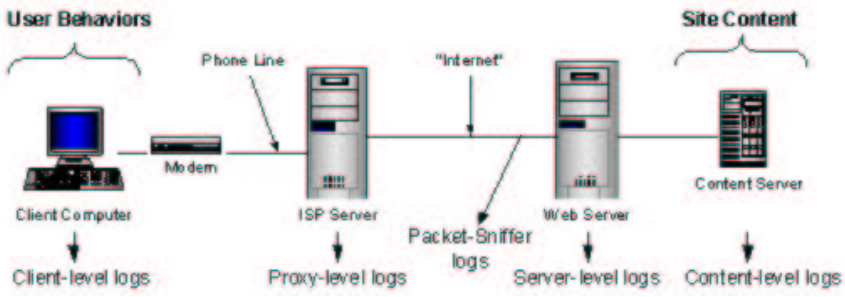


Figure 3.12: High Level Architecture of Different Web Services

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 [LDEKST2002].

3.5.6 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 [DG2000] and [DCLKa] seem to indicate that people value their privacy, while experience at major e-commerce portals shows that over 97can be provided based on it. Spiekerman et al [SGB2001] 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 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.

3.6 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, the prominent successful applications, and outlined some promising areas of future research. Our hope is that this overview provides a starting point for fruitful discussion.

3.7 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, and Pang-Ning Tan, 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.