AHPCRC Technical Report 2003-110

AHPCRC TECHNICAL REPORT

CLARK ATLANTA UNIVERSITY • FLORIDA A&M UNIVERSITY • HOWARD UNIVERSITY • JACKSON STATE UNIVERSITY
UNIVERSITY OF MINNESOTA • UNIVERSITY OF NORTH DAKOTA • NETWORK COMPUTING SERVICES, INC.

WEB MINING - CONCEPTS, APPLICATIONS AND RESEARCH DIRECTIONS

BY

J. SRIVASTAVA, P. DESIKAN AND V. KUMAR

TECHNICAL REPORT NUMBER

2003-110

ARMY HIGH PERFORMANCE COMPUTING RESEARCH CENTER
PO BOX 581459
MINNEAPOLIS, MINNESOTA 55458-1459
http://www.arc.umn.edu/

WEB MINING – CONCEPTS, APPLICATIONS AND RESEARCH DIRECTIONS

BY

J. SRIVASTAVA, P. DESIKAN AND V. KUMAR

Submitted to:

TECHNICAL REPORT NUMBER 2003-110

Supported in part by the United States Army

Cooperative Agreement Number: DAAD19-01-2-0014
Contract Number: DASW01-01-C-0015
University of Minnesota

July 2003

The content of the information contained in this manuscript does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

Approved for Public Release; Distribution Unlimited

Chapter 3

Web Mining - Concepts, Applications & Research 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 history, both amongst researchers and practitioners. This chapter provides a brief overview of the accomplishments of the field, both in terms of technologies and applications, and outlines key future research directions.

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 [69] 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 certainly 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 [40,41,49,50], and the Web Analytics workshops, which have been held in conjunction with the SIAM data mining conference [27,28]. A good survey of the research in the field till the end of 1999 is provided by Kosala and Blockeel [42] and Madria et al. [48].

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 [25]. 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 [18]. The second definition has become more acceptable, as is evident from the approach adopted in most recent papers [8,42,48] that have addressed the issue. In this paper we follow the data-centric view of Web mining which is defined as follows,

> **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 organization, has led to the accumulation of significant experience. It is our goal 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 3.2 we provide a taxonomy of Web mining, in Section 3.3 we summarize some of the key concepts in the field, and in Section 3.4 we describe successful applications of Web mining. In Section 3.5 we present some directions for future research, and in Section 3.6 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. Figure 3.1 shows the taxonomy.

Web Content Mining: Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data is the collection of facts a Web page is designed to contain. 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 include topic discovery and tracking, 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.

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 a 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 Desikan et al. [23] provide 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 [54, 73].

Web Usage Mining: Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web usage data, in order to understand and better serve the needs of Web-based applications [68]. 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*: User logs are collected by the Web server and typically include IP address, page reference and access time.

- *Application Server Data*: Commercial application servers such as Weblogic [6], [10], StoryServer [72] have significant features 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*: New kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these events.

It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories.

3.3 KEY CONCEPTS

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

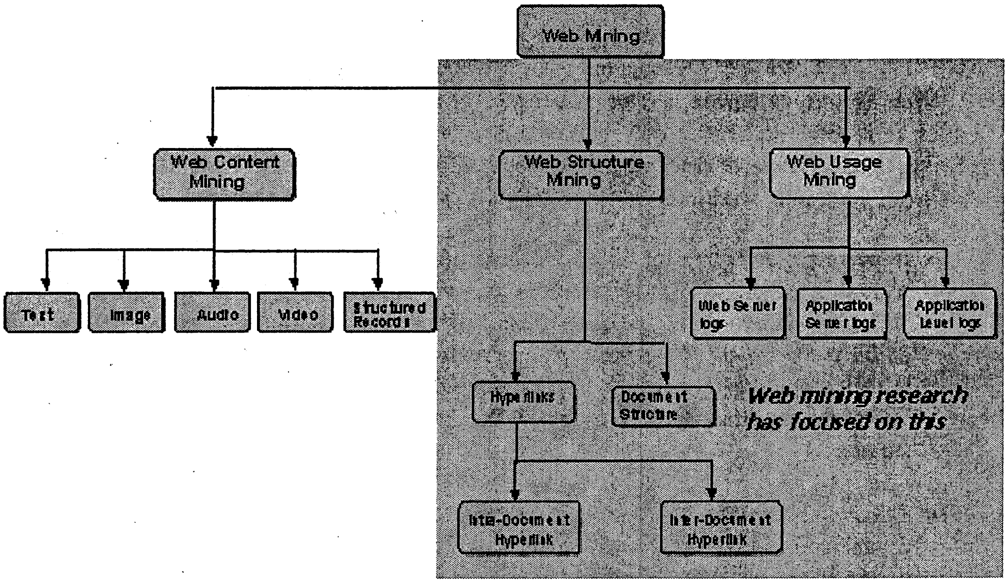

Figure 3.1: Web mining Taxonomy

3.3.1 Ranking metrics - for page quality and relevance.

Searching the Web involves two main steps: *Extracting the pages relevant 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 ones.

PageRank: PageRank is a metric for ranking hypertext documents based on their quality. Page et al. [58] developed this metric for the popular search engine Google [9, 32]. 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 pages are determined. The rank of a page p can be written as:

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

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 is 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 the surfer chooses a URL directly, rather than traversing a link[1] and

[1] The parameter d, called the dampening factor, is usually set between *0.1* and *0.2* [9].

$1 - d$ 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 is the probability of arriving at a page by traversing a link. 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, where the 'fans' represent the hubs and the 'centers' 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 scores are computed for a set of pages related to a topic using an iterative procedure called HITS [38]. 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* [47], *Topic Sensitive PageRank* [33] and *Web page Reputations* [51]. These different hyperlink based metrics have been discussed by Desikan et al. [23].

3.3.2 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 separating robot behavior from human behavior prior to building user behavior models has been illustrated by Kohavi [39]. First, e-commerce retailers are particularly concerned about the unauthorized deployment of robots for gathering business intelligence at their Web sites. Second, 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 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. Tan and Kumar [70] proposed a classification based 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.

3.3.3 Information scent - Applying foraging theory to browsing behavior

Information scent is a concept that uses the snippets of information present 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 [12]. 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 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 then decided by the 'scent' or the value of the element in the 'Scent' matrix.

3.3.4 User profiles - Understanding how users behave

The Web has taken user profiling to 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 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 purposes [50].

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 [3] and DoubleClick [20]. These approaches require browser cookies of some sort, and can provide a fairly detailed view of a user's browsing behavior across the Web.

3.3.5 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 click-stream of a reader to observe the exact path taken through on-line published material. We can measure 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 by Perkowitz and Etzioni [60].

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 so 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 by several authors [16,57].

3.3.6 Pre-processing - making Web data suitable for mining

In the panel discussion referred to earlier [69], 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 [17] , session creation [17] , robot detection and filtering [70] , extracting usage path patterns [66], etc. Cooley's Ph.D. thesis [16] 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 [9]. An up-to-date survey of structure preprocessing is provided by Desikan et al. [23].

3.3.7 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. Gibson et al [29] identified Web communities as "*a core of central 'authoritative' pages linked together by 'hub' pages*".Their approach was extended by Ravi Kumar et al [43] to discover emerging Web communities while crawling. A different approach to this problem was taken by Flake et al [26] who applied the "maximum-flow minimum cut model" [36] to the Web graph for identifying "Web communities". Imafuji et al [34] compare HITS and the maximum flow approaches and discuss the strengths and weakness of the two methods. Reddy et al [62] 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 [2]. 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. Similarity is measured using features such as text, out-links, in-links and mailing lists.

3.3.8 Online Bibiliometrics

With the Web having become the fastest growing and most up to date source of information, the research community has found it extremely useful to have online repositories of publications. Lawrence et al have observed [44] that having articles online makes them more easily accessible and hence more often cited than articles that are 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.

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 [64], ACM portal [1], CiteSeer [14] and DBLP [19].

3.3.9 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 [13] 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 [11] present a tool called WebCANVAS that displays clusters of users with similar navigation behavior. Prasetyo et al [61] developed 'Naviz', an 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. Such a tool is very useful in analyzing user behavior and improving web sites.

3.4 PROMINENT APPLICATIONS

Excitement about the Web in the past few years has led to the Web applications being 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.

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."[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' [55]. 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. [4].

3.4.2 Web Search - Google

Google [32] 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 information from the web. PageRank, which 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 are 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 program 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 to five times.

[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.

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 [30].

One of the latest services offered by Google is,'Google News' [31]. 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 distillation [67].

3.4.3 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 [20]. 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 DoubleClick'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 [3] 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 [21,22]. Microsoft's 'Passport' technology also falls into this category [52]. The value of this technology in applications such as 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 information to law enforcement agencies.

3.4.4 Understanding Web communities - AOL

One of the biggest successes of America Online (AOL) has been its sizeable and loyal customer base [5]. 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.

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 reach, with the convenience of doing it from one's home PC [24]. 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 [15]. Recent efforts are geared towards understanding participants' bidding behaviors/patterns to create a more efficient auction market.

3.4.6 Personalized Portal for the Web - MyYahoo

Yahoo [75] 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 [76] for private information, e.g bank and brokerage accounts. 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]

3.4.7 CiteSeer - Digital Library and Autonomous Citation Indexing

NEC ResearchIndex, also known as CiteSeer [7, 14], is one of the most popular online bibiliographic indices related to Computer Science. The key contribution of the CiteSeer repository is its 'Autonomous Citation Indexing' (ACI) [45]. Citation indexing makes 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.

[3] Yahoo has been consistently ranked as one of the top Web properties for a number of years [53].

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

3.5 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 [74]. 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 provide 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 this revolutionary experimental apparatus for human behavior studies. The Web Lab of Amazon [4] 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, influencing the process, etc. [56]. Underhill [71] has conclusively proven the value of process information in

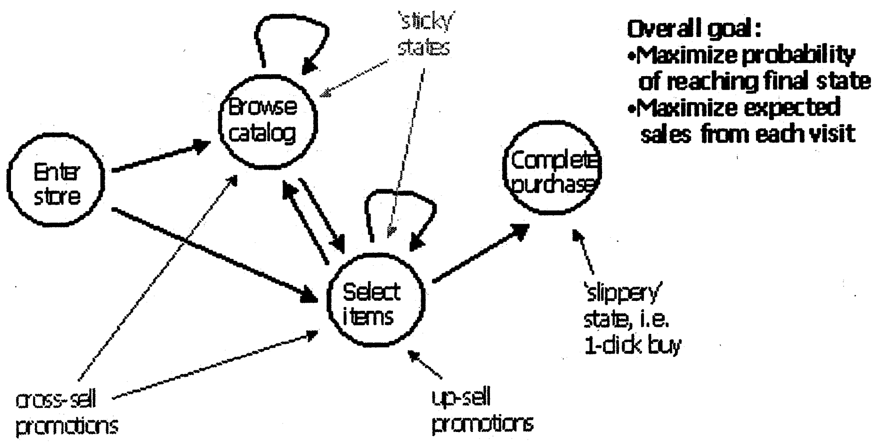

Figure 3.2: Shopping Pipeline modeled as State Transition Diagram

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 3.2 shows an approach of modeling online shopping as a state transition diagram.

3.5.3 Temporal evolution of the Web

Society's interaction with the Web is changing the Web as well as the way people interact with each other. While storing the history 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 [35]. 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 usage data from their 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.

3.5.4 Web services performance optimization

As services over the Web continue to grow [37], 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 [59]. It is necessary to do analysis of the Web logs for web services performance optimization as shown in Figure 3.3. Research is needed in developing Web mining techniques to improve various other aspects of Web services.

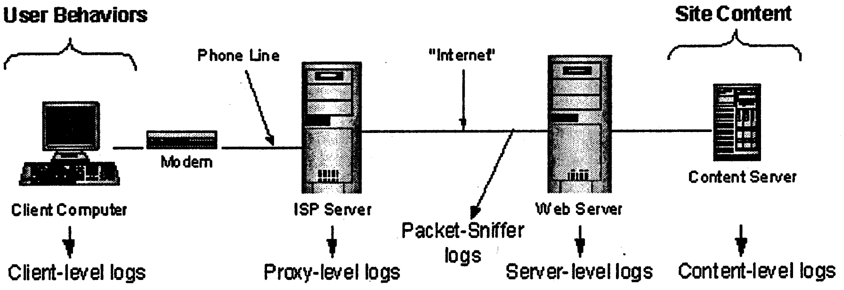

Figure 3.3: High Level Architecture of Different Web Logs

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 [63]. Yet another example is auction fraud, which has been increasing on popular sites like eBay. 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, characterize them and recognize emerging frauds. The issues in cyber threat analysis and intrusion detection are quite similar in nature [46].

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 cases like [22] and [21] 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 are provided based on it. Spiekerman et al [65] 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, and 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 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, a number of prominent applications, and outlined some 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, 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.

Bibliography

[1] ACM Portal. http://portal.acm.org/portal.cfm.

[2] L. Adamic and E. Adar. Friends and Neighbors on the Web. Xerox, Paolo Alto Research Center, CA.

[3] Alexa research. http://www.alexa.com.

[4] Amazon.com. http://www.amazon.com.

[5] America Online. http://www.aol.com, 2002.

[6] BEA Weblogic Server. http://www.bea.com/products/weblogic/server/index.shtml.

[7] 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.

[8] J. Borges and M. Levene. Mining Association Rules in Hypertext Databases. In *Knowledge Discovery and Data Mining*, pages 149–153, 1998.

[9] 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.

[10] Broadvision 1-to-1 portal. http://www.bvportal.com/.

[11] 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.

[12] 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.

[13] 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.

[14] CiteSeer Scientific Literature Digital Library. `http://citeseer.nj.nec.com/cs`.

[15] E. Colet. Using Data Mining to Detect Fraud in Auctions, 2002.

[16] R. Cooley. *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*. PhD thesis, University of Minnesota, 2000.

[17] 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.

[18] R. Cooley, J. Srivastava, and B. Mobasher. Web mining: Information and pattern discovery on the world wide web. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, 1997.

[19] DBLP Bibiliography. `http://www.informatik.uni-trier.de/~ley/db/`.

[20] DoubleClick's DART Technology. `http://www.doubleclick.com/dartinfo/`, 2002.

[21] DoubleClick's Lawsuit. `http://www.wired.com/news/business/0,1367,36434,00.html`, 2002.

[22] C. Dembeck and P. A. Greenberg. Amazon: Caught Between a Rock and a Hard Place. `http://www.ecommercetimes.com/perl/story/2467.html`, 2002.

[23] P. Desikan, J. Srivastava, V. Kumar, and P.N. Tan. Hyperlink AnalysisTechniques & Applications. Technical Report 2002-152, Army High Performance Computing Research Center, 2002.

[24] eBay Inc. `http://www.ebay.com`.

[25] O. Etzioni. The World-Wide Web: Quagmire or Gold Mine? *Communications of the ACM*, 39(11):65–68, 1996.

[26] 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.

[27] J. Ghosh and J. Srivastava. Proceedings of Workshop on Web Analytics. `http://www.lans.ece.utexas.edu/workshop\_index2.htm`, 2001.

[28] J. Ghosh and J. Srivastava. Proceedings of Workshop on Web Mining. `http://www.lans.ece.utexas.edu/workshop\_index.htm`, 2001.

[29] D. Gibson, J.M. Kleinberg, and P. Raghavan. Inferring Web Communities from Link Topology. In *UK Conference on Hypertext*, pages 225–234, 1998.

[30] Google Recognized As Top Business-To-Business Media Property. `http://www.google.com/press/pressrel/b2b.html`.

[31] Google News. `http://news.google.com`.

[32] Google Inc. `http://www.google.com`.

[33] T. Haveliwala. Topic-sensitive PageRank. In *In Proceedings of the Eleventh International World Wide Web Conference, Honolulu, Hawaii, May 2002.*, 2002.

[34] 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.

[35] The Internet Archive Project. `http://www.archive.org/`.

[36] L.R. Ford Jr and D.R. Fulkerson. Maximal Flow through a network. *Canadian J. Math*, 8:399–404, 1956.

[37] R.H. Katz. Pervasive Computing: It's All About Network Services, 2002.

[38] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[39] 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.

[40] R. Kohavi, B. Masand, M. Spiliopoulou, and J. Srivastava. Proceedings of WebKDD2001 - Mining Log Data Across All Customer Touchpoints, 2001.

[41] R. Kohavi, M. Spiliopoulou, and J. Srivastava. Proceedings of WebKDD2000 - Web Mining for E-Commerce - Challenges and Opportunities, 2001.

[42] 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.

[43] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1481–1493, 1999.

[44] S. Lawrence. Online or invisible? *Nature*, 411(6837):521, 2001.

[45] S. Lawrence, C.L. Giles, and K. Bollacker. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.

[46] A. Lazarevic, P. Dokas, L. Ertoz, V. Kumar, J. Srivastava, and P.N. Tan. Data mining for network intrusion detection. In *NSF Workshop on Next Generation Data Mining*, 2002.

[47] 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.

[48] 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.

[49] B. Masand and M. Spiliopoulou. Proceedings of WebKDD1999 - Workshop on Web Usage Analysis and User Profiling, 1999.

[50] B. Masand, M. Spiliopoulou, J. Srivastava, and O. Zaiane. Proceedings of WebKDD2002 - Workshop on Web Usage Patterns and User Profiling, 2002.

[51] A.O. Mendelzon and D. Rafiei. What do the Neighbours Think? Computing Web Page Reputations. *IEEE Data Engineering Bulletin*, 23(3):9–16, 2000.

[52] MicroSoft.NET Passport. `http://www.microsoft.com/netservices/passport/`.

[53] Top 50 US Web and Digital Properties. `http://www.jmm.com/xp/jmm/press/mediaMetrixTop50.xml`.

[54] C.H. Moh, E.P. Lim, and W.K. Ng. DTD-Miner: A Tool for Mining DTD from XML Documents. WECWIS, 2000.

[55] E. Morphy. Amazon pushes 'personalized store for every customer'. `http://www.ecommercetimes.com/perl/story/13821.html`, 2001.

[56] K.L. Ong and W. Keong. Mining Relationship Graphs for Eective Business Objectives.

[57] B. Padmanabhan and A. Tuzhilin. A Belief-Driven Method for Discovering Unexpected Patterns. In *Knowledge Discovery and Data Mining*, pages 94–100, 1998.

[58] 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.

[59] A. Pandey, J. Srivastava, and S. Shekhar. A web intelligent prefetcher for dynamic pages using association rules - a summary of results, 2001.

[60] M. Perkowitz and O. Etzioni. Adaptive Web Sites: Conceptual Cluster Mining. In *IJCAI*, pages 264–269, 1999.

[61] B. Prasetyo, I. Pramudiono, K. Takahashi, M. Toyoda, and M. Kitsuregawa. Naviz user behavior visualization of dynamic page.

[62] 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.

[63] D. Scarponi. Blackmailer Reveals Stolen Internet Credit Card Data. http://abcnews.go.com/sections/world/DailyNews/internet000110.html, 2000.

[64] Science Citation Index. http://www.isinet.com/isi/products/citation/sci/.

[65] 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.

[66] M. Spiliopoulou. Data Mining for the Web. Proceedings of the Symposium on Principles of Knowledge Discovery in Databases (PKDD), 1999.

[67] T. Springer. Google LaunchesNews Service. http://www.computerworld.com/developmenttopics/websitemgmt/story/0,1080%1,74470,00.html, 2002.

[68] 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.

[69] J. Srivastava and B. Mobasher. Web Mining: Hype or Reality? . 9th IEEE International Conference on Tools With Artificial Intelligence (ICTAI '97), 1997.

[70] P. Tan and V. Kumar. Discovery of web robot sessions based on their navigational patterns. *Data Mining and Knowledge Discovery*, 2002.

[71] P. Underhill. Why we buy: The Science of shopping. Touchstone Books, 2000.

[72] Vignette StoryServer. http://www.cio.com/sponsors/110199_vignette_story2.html.

[73] 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.

[74] Hosting Firm Reports Continued Growth. http://thewhir.com/marketwatch/ser053102.cfm, 2002.

[75] Yahoo!, Inc. http://www.yahoo.com.

[76] Yodlee, Inc. http://www.yodlee.com.

AHPCRC Available Technical Reports

The Army High Performance Computing Research Center provides the opportunity for interested parties to keep abreast of the latest research being done in the Center through the AHPCRC Technical Report Series. This Series consists of papers written by participating researchers and their colleagues. Each technical report is duplicated and bound for convenient referencing and is available upon request.

A WWW database on the AHPCRC and its research activities can be accessed via any WWW browser. To view this database, open the following URL: "http://www.arc.umn.edu/". For additional assistance, send email to Shelli Quackenboss (shelli@cs.umn.edu). Current reports in the series include:

2003-110 J. Srivastava, P. Desikan and V. Kumar
Web Mining – Concepts, Applications and Research

2003-109 Bidhan C. Saha
Slow Li + He Collisions: A Molecular State Treatment

2003-108 L. Ertoz, E. Eilertson, A. Lazarevic, P-Ning Tan, P. Dokas, V. Kumar and J. Srivastava
Detection and Summarization of Novel Network Attacks Using Data Mining

2003-107 M.G.A. Tijssens, R.D. James
Towards An Improved Continuum Theory For Phase Transformations

2003-106 G. Karypis
Multi-Constraint Mesh Partitioning for Contact/Impact Computations

2003-105 B. Chen and H. Xu
ActivePoints: Acquisition, Processing and Navigation of Large Outdoor Environments

2003-104 B. Chen and H. Xu
ActivePoints: Acquisition, Processing and Navigation of Large Outdoor Environments

2003-103 B. Chen and H. Xu
Stylized Visualization of 3D Scanned Outdoor Environments

2003-102 L. Ertoz, M. Steinbach and V. Kumar
Finding Clusters of Different Sizes, Shapes and Densities in Noisy, High Dimensional Data

2003-101 A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur and J. Srivastava
A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection

2003-100 B. Chen and H. Xu
ActivePoints: Acquisition, Processing and Navigation of Large Outdoor Environments

2002-173 J. Leszczynski, A. Michalkova and L. Gorb
Interactions of Model Organic Species and Explosives with Clay Minerals

2002-172 J. Leszczynski
Exploring Clean-Up Technologies For Explosives in Soils and Groundwater

2002-171 H. Xiong, P.-Ning Tan, and V. Kumar
Mining Hyperclique Patterns with Confidence Pruning

2002-170 B. Chen, F. Dachille and A. Kaufman
Footprint Area Sampled Texturing

2002-169 B. Chen and G. Dahl
Digital Synthesis Control

2002-168 B. Chen, X. Yuan, M. Nguyen and H. Xu
Hybrid Forward Resampling and Volume Rendering

2002-167 S. Aliabadi and S. Zhang Tu
A Discontinuous Galerkin Method with Limiter for Advection-Diffusion Problems and its Extension for Compressible Flows

2002-166 S. Aliabadi, A. Johnson, B. Zellars and J. Abedi
Parallel Simulation of Waves interacting with Ships in Motion

2002-165 J. Srivastava, P. Desikan, and V. Kumar
Web Mining – Accomplishments and Future Directions

2002-164 J. Srivastava and R. Cooley
Web Business Intelligence: Mining the Web for Actionable Knowledge

2002-163 J. Srivastava, J. Wang, E. Lim, and S. Hwang
A Case for Analytical Customer Relationship Management

2002-162 R. Zalesny, W. Bartkowiak, S. Styrcz, and J. Leszczynski
Solvent Effects on Conformationally Induced Enhancement of the Two-Photon Absorption Cross Section of a Pyridinium-N-Phenolate Betaine Dye. A Quantum Chemical Study