Web Mining for Business Computing

Prasanna Desikan, Colin DeLong, Sandeep Mane, Kalyan Beemanapalli, Kuo-Wei Hsu, Prasad Sriram, Jaideep Srivastava, Vamsee Venuturumilli
University of Minnesota

Abstract

Over the last decade, there has been a paradigm shift in business computing with the emphasis moving from data collection to knowledge extraction. Central to this shift has been the explosive growth of the World Wide Web, which has enabled myriad technologies, such as Web services and enterprise server applications. These advances have improved data collection frameworks and resulted in new techniques for knowledge extraction from large databases. A popular and successful technique which has showed much promise is Web mining. Web mining is essentially data mining for Web data, thus enabling businesses to turn their vast repositories of transactional and Website usage data into actionable knowledge that is useful at every level of the enterprise – not just the front-end of an online store. To this end, the chapter provides an introduction to the field of Web mining and examines existing as well as potential Web mining applications applicable for different business function, like marketing, human resources, and fiscal administration. Suggestions for improving information technology infrastructure are made, which can help businesses interested in Web mining hit the ground running.

1 Introduction

The Internet has changed the rules for today's businesses, which now increasingly face the challenge of improving and sustaining performance throughout the enterprise. The growth of the World Wide Web and enabling technologies has made data collection, data exchange and information exchange easier and has resulted in speeding up of most major business functions. Delays in retail, manufacturing, shipping, and customer service processes are no longer accepted as necessary evils, and firms improving upon these (and other) critical functions have an edge in their battle of margins. Technology has been brought to bear on myriad business processes and affected massive change in the form of automation, tracking, and communications, but many of the most profound changes are yet to come.

Leaps in computational power have enabled businesses to collect and process large amounts of data. The availability of data and the necessary computational resources, together with the potential of data mining, has shown great promise in having a transformational effect on the way businesses perform their work. Well-known successes of companies such as Amazon.com have provided evidence to that end. By leveraging large repositories of data collected by corporations, data mining techniques and methods offer unprecedented opportunities in understanding business processes and in predicting future behaviour. With the Web serving as the realm of many of today's businesses, firms can improve their ability to know when and what customers want by understanding customer behaviour, find bottlenecks in internal processes, and better anticipate industry trends.

This chapter examines past success stories, the current efforts, and future directions of 'Web mining' as an application for business computing. Examples are given in different business aspects, such as product recommendations, fraud detection, process mining, inventory management, and how the use of Web mining will enable growth revenue, minimize costs, and enhance strategic vision. Gaps in existing technology are also explained, along with pointers to future directions.

2 Web Mining

Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, and usage logs of Web sites. A panel organized at ICTAI 1997 (Srivastava and Mobasher, 97) 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 past decade, and the number of significant ideas that have been developed have answered this question in the affirmative. In addition, a fairly stable community of researchers interested in the area has been formed, through the successful series of workshops such as WebKDD (held annually in conjunction with the ACM SIGKDD Conference) and the Web Analytics (held in conjunction with the SIAM data mining conference). Many informative surveys exist in the literature that addresses various aspects of Web mining (Cooley et al, 1997; Kosala and Blockeel, 2000; Mobasher, 2005).

Two different approaches have been taken in defining Web mining. First was a 'process-centric view', which defined Web mining as a sequence of tasks (Etzioni, 1996). 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 (Cooley et al, 1997). The second definition has become more acceptable, as is evident from the approach adopted in most recent papers that have addressed the issue. In this chapter, we use the data-centric view of Web mining, which is defined as,

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

The attention paid to Web mining in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences. Its application in business computing has also found tremendous utility. In the following sub-sections, we describe the taxonomy of Web mining research and applicability of Web mining to business computing.

2.1 Web Mining Taxonomy

Web mining can be broadly divided into three distinct categories according to the kinds of data to be mined. We provide a brief overview of the three categories and an illustration depicting the taxonomy is shown in Figure 2.

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 information on a Web page, which is conveyed to users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining include topic discovery, extracting association patterns, clustering of Web documents and classification of Web Pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While a significant body of work in extracting knowledge from images, in the fields of image processing and computer vision exists, the application of these techniques to Web content mining has been limited.

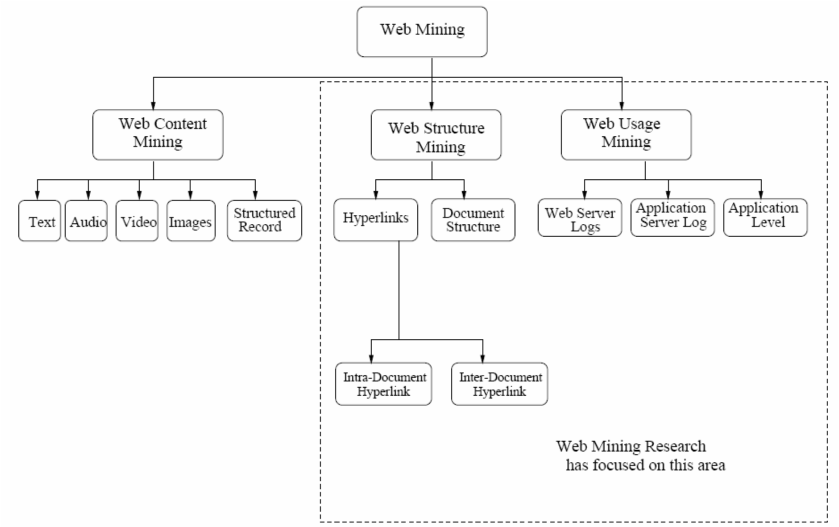


Figure 1: Web Mining Taxonomy

Web structure mining: Web structure mining is the process of discovering structure information from the Web. The structure of a typical Web graph consists of Web pages as nodes and hyperlinks as edges connecting related pages. Web structure mining can be further divided into two kinds based on the type of structured information used.

- **Hyperlinks:** A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an *Intra-Document Hyperlink*, and a hyperlink that connects two different pages is called an *Inter-Document Hyperlink*. There has been a significant body of work on hyperlink analysis (see survey paper on hyperlink analysis, Desikan et al, 2002).

- **Document Structure:** 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. Here, mining efforts have focused on automatically extracting document object model (DOM) structures out of documents.

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. Usage data captures the identity or origin of Web users along with their browsing behaviour at a Web site. Web usage mining itself is further classified depending on the kind of usage data used:

- **Web Server Data:** The user logs are collected by Web server. Typical data includes IP address, page reference and access time.

- **Application Server Data:** Commercial application servers, e.g. Weblogic, 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:** 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.

2.2 Web mining research – state of the art

The interest of research community and the rapid growth of work in this area have resulted in significant research contributions which has been summarized in a number of surveys and book chapters over the past few years (Kosala and Blockeel, 2000; Cooley et al, 1997; Srivastava et al, 2004). Research on Web content mining has focused on issues such as extracting information from structured and unstructured data and integrating information from various sources of content. Earlier work on Web content mining can be found in Kosala's work (Kosala and Blockeel, 2000). Web content mining together has found its utility in a variety of applications such as Web page categorization and topic distillation. A special issue on Web content mining (Liu and Chang, 2004) captures the recent issues that have drawn the attention of the research community in Web content mining. Web structure mining has focused primarily on hyperlink analysis. A survey on hyperlink analysis techniques and a methodology to pursue research has been proposed by Desikan et al (Desikan et al 2002). Most of these techniques can be used independently or in conjunction with techniques proposed with Web content and Web usage. The most popular application is ranking of Web pages. PageRank (Page et al, 1998), developed by Google founders, is a popular metric for ranking the importance of hypertext documents for Web search. The key idea in PageRank is that a page has a high rank if many highly ranked pages point it to, and hence the rank of a page depends upon the ranks of pages pointing to it. Another popular measure is *hub* and *authority* scores. The underlying model for computing these scores is a bipartite graph (Kleinberg, 1998). The Web pages are modeled as 'fans' and 'centers' of a bipartite core, where a 'fan' is regarded as a hub page and 'center' as an authority page. For a given query, a set of relevant pages is retrieved. And for each page in such a set, a hub score and an authority score (Kleinberg, 1998).

Web usage data has is the key to understand user's perspective of the Web, while content and structure reflect the creator's perspective. Understanding user profiles and user navigation patterns for better adaptive Web sites and predicting user access patterns has evoked interest to the research and the business community. The primary step for Web usage mining is pre-processing the user log data, such as to separate Web page references into those made for navigational purposes and those made for content purposes (Cooley et al, 1999). The concept of adaptive Web was introduced by researchers from University of Washington, Seattle (Perkowitz and Etzioni, 97). Markov models have been the most popular form of techniques to predict user behaviour (Pirolli and

Pikow, 99; Sarukkai, 1999; Zhu et al, 2002). A more detailed information about various aspects of Web usage mining techniques can be found in a recent extensive survey on this topic (Mobasher, 2005).

3 How Web Mining Can Enhance Major Business Functions

This section discusses existing and potential efforts in the application of Web mining techniques to the major functional areas of businesses. Some examples of deployed systems as well as frameworks for emerging applications yet-to-be-built are discussed. However, the examples are no means to be regarded as solutions to all problems within the framework of business function they are cited in. Their purpose is to illustrate that Web mining techniques have been applied successfully to handle certain kind of problems, providing the evidence of its utility. Table 1 at the end of this section (page 27) provides the summary of how Web mining techniques have been successfully applied to address various issues that arise in business functions.

3.1 Marketing

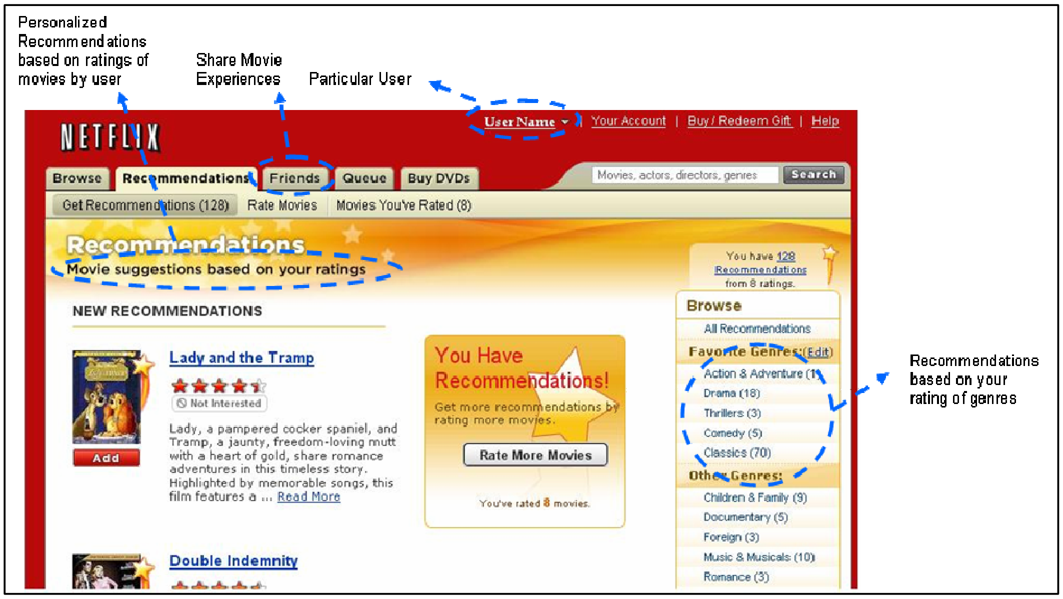
Marketing is typically defined (Brian Norris, 2005) as:

> "Marketing is the ongoing process of moving people closer to making a decision to purchase, use, follow or conform to someone else's products, services or values. Simply, if it doesn't facilitate a 'sale', then it's not marketing"

Marketing is responsible for keeping the enterprise attentive to market trends, as well as keeping the sales unit aware of where the target segment is. In the following examples, we illustrate how Web mining techniques have been used for marketing products to a customer and also to identify possible new areas of potential market for an enterprise.

Product recommendation

Recommending products to purchase is a key issue for all businesses. As the customer-centric approach drives the current business models, traditional brick-and-mortar stores have to rely on data collected explicitly from customers through surveys to offer customer-centric recommendations. However, the advent of e-commerce not only enables a level of closeness in customer-to-store interaction (that is far greater than imaginable in the physical world), but also leads to unprecedented data collection, especially about the ‘process of shopping’. The desire to understand individual customer’s shopping behaviour and psychology in detail, by mining that data has led to significant advances in on-line customer relationship management (e-CRM) and providing services like real-time recommendations. A recent survey (Adomavicius and Tuzhilin, 2005) provides an excellent taxonomy of various techniques that have been developed for on-line recommendations.



Personalized Recommendations based on ratings of movies by user

Share Movie Experiences

Particular User

Recommendations based on your rating of genres

Lady, a pampered cocker spaniel, and Tramp, a jaunty, freedom-loving mutt with a heart of gold, share romance adventures in this timeless story. Highlighted by memorable songs, this film features a ...

Figure 2: NetFlix.com - an example of product recommendation using Web usage mining

NetFlix.com[1] is a good example of how an online-store uses Web mining techniques for recommending products, such as movies, to customers based on their past rental profile and profiles of users who have similar

[1] It should be noted that NetFlix.com was chosen as an example of an online store with no particular preference shown over other online stores.

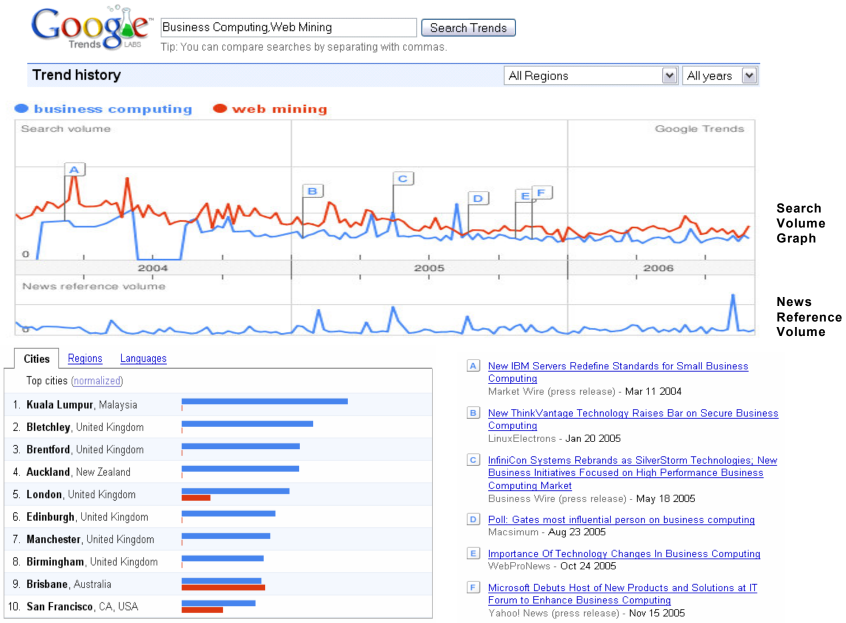
browsing and renting patterns. A host of Web mining techniques, such as associations between pages visited and click-path analysis are used to improve the customer's experience and provide recommendations during a "store visit." Techniques for automatic generation of personalized product recommendations (Mobasher et al 2000) form the basis of most recommendation models. Knowledge gained from Web mining is the key intelligence behind NetFlix's features such as favourite genres, recommendation based on earlier movies rated by user, or recommending based on shared information with friends, who are a part of their social network.

Product Area and Trend Analysis

John Ralston Saul, the Canadian author, essayist and philosopher noted:

> "*With the past, we can see trajectories into the future - both catastrophic and creative projections*."

Businesses would definitely like to see such projections onto the future. Specially, identifying new product areas based on trends is a key for any business to capture markets. Prediction using trend analysis for a new product usually addresses two kinds of issues. Firstly, the potential market for a particular product. Secondly, a single product may result in a platform to develop a class of products that have a high potential market. Different methods have been implemented for such prediction purposes. Among the popular techniques are surveying techniques and time-series forecasting techniques. Traditionally, sufficient data collection was a major hurdle in the application of such techniques. However, with the advent of the Web, the hurdle of filling up forms and communicating has reduced to a simple series of clicks. This enabling technology has caused a huge shift in the amount of data collected, and more significantly in understanding the customer behaviour. For example, applying Web mining to the data collected from online community interactions provides a very good understanding of its communities, which is then used for targeted marketing through advertisements and e-mail solicitation. A good example is AOL's 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 that 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 existing and new products, and also test strategies for influencing opinions. New product sales can also be modelled using other techniques, such as co-integration analysis (Franses 1994).



Top cities (normalized)

1. **Kuala Lumpur**, Malaysia
2. **Bletchley**, United Kingdom
3. **Brentford**, United Kingdom
4. **Auckland**, New Zealand
5. **London**, United Kingdom
6. **Edinburgh**, United Kingdom
7. **Manchester**, United Kingdom
8. **Birmingham**, United Kingdom
9. **Brisbane**, Australia
10. **San Francisco**, CA, USA

A New IBM Servers Redefine Standards for Small Business Computing
Market Wire (press release) - Mar 11 2004

B New ThinkVantage Technology Raises Bar on Secure Business Computing
LinuxElectrons - Jan 20 2005

C InfiniCon Systems Rebrands as SilverStorm Technologies; New Business Initiatives Focused on High Performance Business Computing Market
Business Wire (press release) - May 18 2005

D Poll: Gates most influential person on business computing
Macsimum - Aug 23 2005

E Importance Of Technology Changes In Business Computing
WebProNews - Oct 24 2005

F Microsoft Debuts Host of New Products and Solutions at IT Forum to Enhance Business Computing
Yahoo! News (press release) - Nov 15 2005

Figure 3: Google Trends

The second most popular technique is time series analysis. Box and Jenkins give an excellent account of various time series analysis and forecasting techniques in their book (Box and Jenkins 94). Time series analysis can also used for decision making in business administration (Arsham 2006). These techniques are generic and can be applied appropriately for predicting trends in product potentials. While most of these techniques have been based on statistical approaches, recent work have shown the data mining can be successfully used to discover patterns of interest in time series data. Keogh (2004) provides a good overview of data mining techniques in time series analysis.

With the advent of Web search and ads based on keyword searches, query words have assumed lot of significance in advertising world. These query words represent popular topics or products among users. Search engines have been focusing on analyzing trends in these query words for improving query-related ads. Figure 4 gives an example of how keywords can be analyzed for their "Trends". The figure depicts the trends in keywords, 'Web Mining' and 'Business Computing'. For example, a possible conclusion seems that the two keywords are having similar trends recently suggesting possible interest in collaboration of the two fields. The news-articles represent randomly selected news articles on a particular topic when the search for topic was high.

3.2 Human Resources

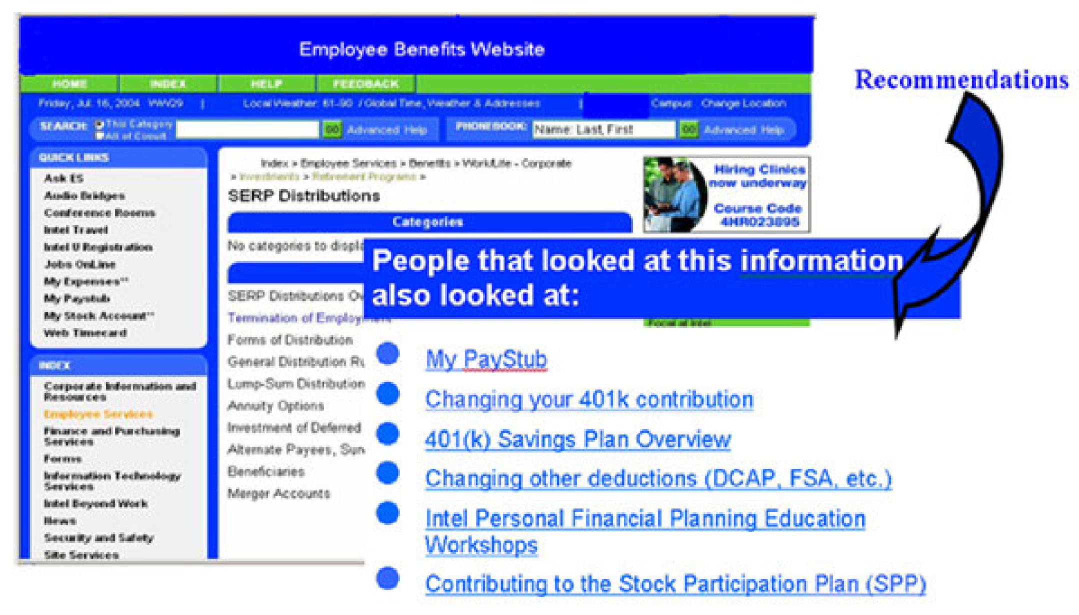
In any enterprise, the broad responsibility of Human Resource department is to correctly match the right skilled personnel with the right function. Human Resources is also responsible to establish policies, guidelines and to provide tools for employees and management to enable a pleasant work atmosphere, strong culture, healthy and save environment, and to ensure that the firm's employees are consistently getting motivated. The following application examines how to effectively manage human resource department by maintaining the right amount of workforce in terms of cost-effectiveness. It illustrates the use of Web mining techniques to reduce unnecessary human workload.

HR Call Centres

Human resource departments of large companies often face the task of answering the numerous questions of various employees. As the size of the company grows and due to globalization, the task becomes more difficult as they not only need to handle the number of employees, but also consider other issues such as geographically local policies and issues. Most of these tasks are significant to the human resource department as it is their duty to keep their employees satisfied and well –informed. A possible and popular approach to handle this problem is to have "call-centres" that provide the informative service to the employees. With the advent of Web, most companies have tried to put all their policy information on Web sites for easy perusal by the employees. However, it has been observed that over time, more and more employees seek the advice of the representatives at the call-centre because of the ambiguity they face while browsing the Website. This phenomenon is termed as "Call Centre Escalation".

The problem of call centre escalation is well-known to customer service centres of commercial products. The problem of expensive call-centres has a resulted in most enterprises having an internal HR Website to serve employees. A recent study (Bose, 2006) talks about real-world enterprise with approximately 80,000 employees and 8000 Web pages for its HR Website experienced the phenomenon of "Call Centre Escalation". Most employees browsing the Website were not able to find relevant information. As a result, employees contacted call centre service for extra information. The problem was not the non-availability of the information, but the difficulty to determine most relevant information within a huge collection of documents.

Web Mining can be used to help the users of Website to find most relevant information for their particular needs. The recent study (Bose, 2006) has shown initial promise in this direction. They have successfully employed Web mining techniques in conjunction with sequence similarity approaches from bioinformatics, to develop a recommendation system that serves the employees with an aim to reduce their click through path length to answers. The technique used, in particular, couples conceptual and structural characteristics of the Website to determine relevant pages for a particular topic. The conceptual characteristics represent the logical organization of Website as designed by Website administrator. The structural characteristics provide a navigational path starting from a particular page. In essence, by using this information, expert knowledge is incorporated into usage mining. Recent studies have shown that these conceptual and structural characteristics of the Website play an important role in improving the predictive power of recommendation engines.



People that looked at this information also looked at:

- My PayStub
- Changing your 401k contribution
- 401(k) Savings Plan Overview
- Changing other deductions (DCAP, FSA, etc.)
- Intel Personal Financial Planning Education Workshops
- Contributing to the Stock Participation Plan (SPP)

Figure 4: Recommendations provided to the user of an Employee benefits Website

Figure 5 gives an example of an Employee benefits Website, with a sample of recommendations provided to a user looking for information related to 401(K) plan.

3.3 Sales Management

In an enterprise, sales is a key responsible function that sells the enterprise's core-competency to customers, usually, with the final goal of bringing in and maximizing revenue from them. As is intuitive, in order to maximize the sales revenue, it is important to both bring in the customers to door, as well as to execute a good sales and operations management strategy, thus resulting in the final sale. Identification of new sales opportunities and the associated risks plays a crucial role in deciding new projects. The following example describes an approach to evaluate the risks of new business opportunities.

Business opportunity risk evaluation

With growing competitive markets, better understanding of customer’s requirements and matching those to the enterprise’s offerings have gained prominence in an enterprise’s decision making processes. Important financial and business forecasts are affected by decisions in such processes and hence these decisions highly influence how an enterprise plans to support its market. For example, a lot of historical data about business sales opportunities are gathered by enterprises for one such analysis. Traditionally, this information of an enterprise’s offerings, competitor’s offerings and the market’s demands are analyzed manually by human experts, using statistical methods, usually using a multi-step process. Correct analysis in such a multi-step process is of prime importance. For example, classifying good (profitable) opportunities as bad (non-profitable) makes the predictions pessimistic and results in lost revenue, whereas classifying bad opportunities (non-profitable) as good opportunities ties up an enterprise’s resources, in addition to asking for unrealistic goals.

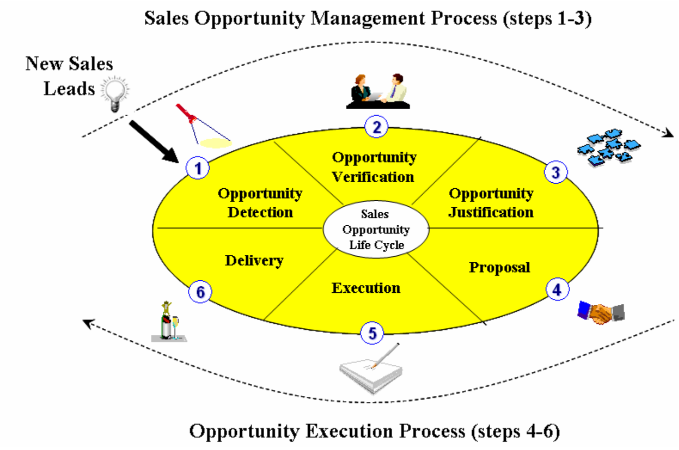
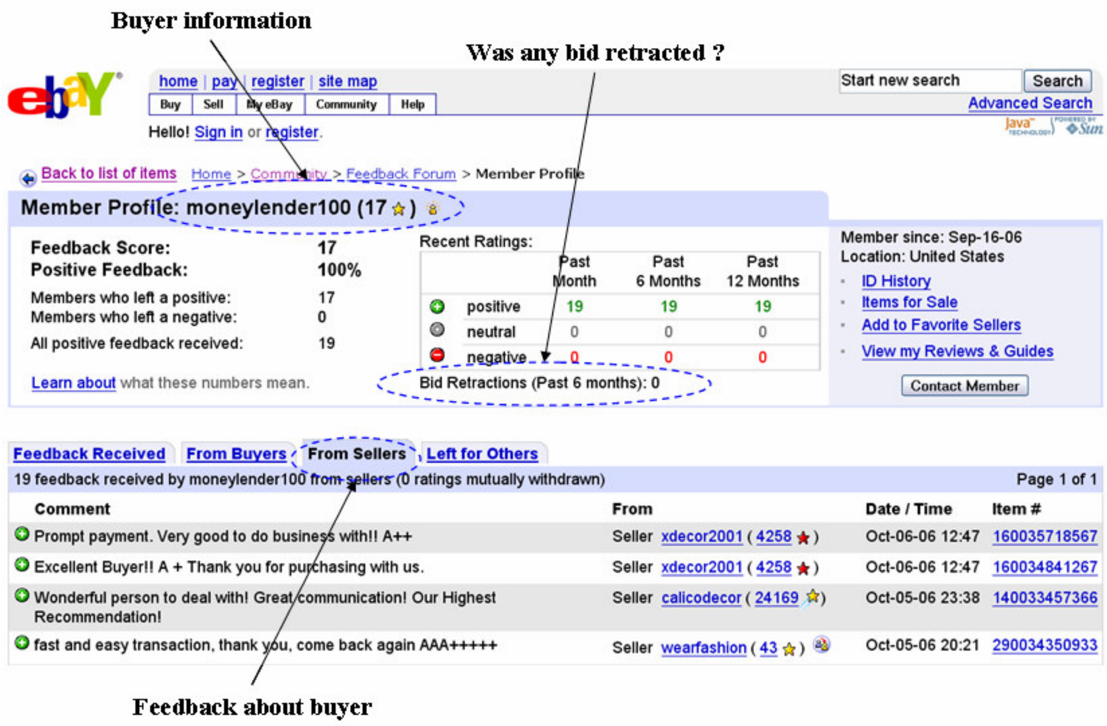

Sales Opportunity Management Process (steps 1-3)
New Sales Leads
1
2
3
4
5
6
Opportunity Detection
Opportunity Verification
Opportunity Justification
Sales Opportunity Life Cycle
Delivery
Execution
Proposal
Opportunity Execution Process (steps 4-6)

Figure 5. Sales opportunity management process (Reference: Mane et al, 2005)

Figure 6 illustrates an example of a multi-step sales opportunity analysis process used in an enterprise, called the “*pipeline analysis*” (Mane et al, 2005), consisting of two sub-processes namely (i) Sales Opportunity Management Process (SOMP), in which sales opportunities are analyzed by managers till the creation of a project, and (ii) Opportunity Execution Process (OEP), involving steps takes for execution of the project. The former sub-process is responsible for the quality and reliability of the sales opportunity data. The accuracy and reliability of pipeline analysis heavily depends on the quality of sales opportunity data and the techniques used to make sales predictions. Manual analysis in such a process is faced with problems like shortage of skilled resources, slower execution speed, and human errors due to an expert's limited ability to extract and use all information in large historical data.



Member Profile: moneylender100 (17)

Feedback Score:	17
Positive Feedback:	100%
Members who left a positive:	17
Members who left a negative:	0
All positive feedback received:	19

Recent Ratings:

	Past Month	Past 6 Months	Past 12 Months
positive	19	19	19
neutral	0	0	0
negative	0	0	0

Bid Retractions (Past 6 months): 0

Member since: Sep-16-06
Location: United States

19 feedback received by moneylender100 from sellers (0 ratings mutually withdrawn) Page 1 of 1

Comment	From	Date / Time	Item #
Prompt payment. Very good to do business with!! A++	Seller xdecor2001 (4258)	Oct-06-06 12:47	160035718567
Excellent Buyer!! A + Thank you for purchasing with us.	Seller xdecor2001 (4258)	Oct-06-06 12:47	160034841267
Wonderful person to deal with! Great communication! Our Highest Recommendation!	Seller calicodecor (24169)	Oct-05-06 23:38	140033457366
fast and easy transaction, thank you, come back again AAA+++++	Seller wearfashion (43)	Oct-05-06 20:21	290034350933

Figure 6: Generation of sales leads

Web and related technologies have enabled collection of such huge amounts of data about business processes. This has led to data mining, and more particularly Web mining, as an emerging area of research interest for business analytics (Apte et al, 2002). Supervised learning techniques can be used to extract knowledge from such data, and thus help experts (e.g. managers) in their analysis. Supervised learning techniques for multi-stage decision processes form an emerging research area. Another problem in such a decision process is a data instance may be split into multiple data instances in intermediate step(s), and/or multiple data instances may be combined into a single data instance. Such link information between data instances across multiple-stages is useful and techniques need to be developed to include this in supervised learning for multi-stage decision process. Link analysis will be useful in developing such new techniques for complex decision processes.

As an extension to previous example, consider the generation of sales leads for the sales opportunity management process. With the emergence of business-to-business, business-to-customer and customer-to-customer e-commerce, Web sites like e-Bay now collect gigabytes of data about buying behaviour of customers. In addition to customer's buying preferences, information about buying power of customers may also be available, Figure 7 illustrates such an example from e-Bay, which allows data, like whether a buyer honors his/her bids, can be collected. Such detailed information was previously difficult to obtain prior to emergence to Web. Matching these data to an enterprise's offerings will help in generation of new sales leads for the enterprise. Such sales leads may then be pursued as sales opportunities in the sales opportunity management process.

3.4 Business Financial Management

Financial Management in an enterprise deals with managing the funds for its different departments and projects with the aim of maximizing profit and minimizing the risk. Financial management itself involves two kinds of issues. The first kind deals with the funds provided to the company from other sources. These funds could be long-term (such as ownership equity) or short-term (such as funding from banks). The second kind deals with 'fund management'. Here, the goals involve identifying issues such as strategies, time lines, and risk-aversion; for the enterprise to make a decision on how much to invest and when. In the following example, we present a case where enabling technologies such as internet and novel techniques such as Web mining can help determine fraud in transactions to help an enterprise reduce its risks.

Fraud analysis

Fraud is a large problem faced by many businesses ranging from the telecommunications industry to Web-based stores. *Fraud* is defined as the use of false representations to gain an unjust advantage or abuse of an organization's resources, such as illegal access to an organization's finances (Bolton and Hand, 2002). For example, credit card fraud causes the loss of millions of dollars to credit card management companies like Visa and MasterCard. This motivates organizations to analyze data in order to identify fraud. However, since large amounts of data are necessary for fraud analysis, it becomes difficult for an organization to manually identify fraud from legitimate transactions. This motivates current research in automated analysis of such data, in order to reduce manual screening of individual transactions for fraudulent activity. There are two approaches to reducing fraud - *fraud prevention*, taking appropriate steps to prevent a fraud from occurring, and *fraud detection*, identifying fraud as soon as it occurs, thus enabling a quick corrective response. Since it is difficult to predict when a fraud has occurred, fraud detection techniques are usually applied in parallel with fraud prevention techniques.

Over the past few years, data mining and machine learning techniques are being applied for such automated fraud detection. The main idea is to compare the observed data with the expected values and then to quantify this deviation as a *suspicion* score, i.e. how likely the observed data belongs to a fraud. Different data mining/statistical approaches have been applied for fraud detection – *supervised learning*, when information about both legitimate instances and frauds is available, *semi-supervised learning*, when information about only legitimate instances is available, and *unsupervised learning*, when no information about whether the data belongs to legitimate instances or fraud is available. Highly skewed class distribution (usually less than 10% of data instances are frauds) and high misclassification costs for classifying frauds as legitimate instances (because of lost revenue) and for classifying legitimate instance as frauds (due to dissatisfied customers) make this a highly challenging task. In addition, the legal definition of frauds may change over time, fraudsters may develop new techniques for frauds, class distribution in data may change, as well as multiple frauds may occur at the same time, thus making fraud detection a highly evolving research field.

Though, historically data mining algorithms were developed for simple attribute-value records, new approaches to determine variable interactions are being investigated. There has been a growing awareness that information about frauds may be spread in the data. This has led to the development of new approaches that would "connect the dots" between such bits of information. For example, in money laundering, a technique called *layering* is used by fraudster(s), wherein a large amount of illegal money is transferred from one set of accounts to another set of accounts using multiple smaller transactions. Link analysis and visualization techniques have been applied for identifying links between such transactions, and thus detecting the fraud. The U.S. Financial Crimes Enforcement Network AI system (FAIS) (Senator et al, 1995) is one such system which allowed the detection of link between transactions, and thus help in detection of money laundering frauds. . With the advent of Web, it is now easier for fraudsters to make such multiple Web-based transactions, while the size of such transactions data has increased several times, thus reducing the chance of being detected. Link analysis, with links updated over time, can be used to establish the "communities of interest" (Cortes et al, 2001) that indicate the networks of fraudsters, thus helping in automated detection of frauds. An emerging research area called *link mining,* studies automated construction of links from a large database based on a particular set of evidence. In addition, techniques are being developed for addressing problem like integrating data from different databases, resolving identities and consolidating data, clustering entities into groups, inferring link information (presence/absence of links and their strength) and constructing (new) unspecified features. Link analysis is also applied for fraud analysis in network intrusion detection, law enforcement, intelligence analysis, and other related domains. However, two problems still faced by fraud detection are the lack of large publicly-available datasets and few publications on current approaches for fraud detection, a reason being that organizations tend to withhold information about which approaches are being used to detect/prevent frauds.

3.5 Production: Shipping and Inventory

Inventory is defined as the value of goods on-hand at a certain time instance, in an enterprise. Inventory can be in different forms like raw material before the firm's value-addition, in process during production phase, as finished product in its warehouse or in a retail-store's distribution center waiting to be sold. Inventory usually results in non-value added handling costs related to tied-up capital, insurance costs, management related cost and also other obsolete inventory costs. Inventory management is defined as a set of activities used to do the right inventory in right place at right time with right quantity and right cost[2]. In the following example application, we discuss how Web mining has aided in inventory management.

Predictive Inventory Management

Since one of significant costs for a business is to maintain an inventory to support sales as well as customers, a successful inventory management helps business decrease cost and increase profit without losing customer satisfaction.

An inventory management system should be able to foresee the customers' demand and trends about sale as well; that is, what most customers will buy and when. Further, in order to help business perform just-in-time inventory, an inventory management system is required to analyze transaction data and accordingly find clusters, each of which is composed of similar items.

As the Internet grew rapidly, people started to get used to Web surfing, leading to emergence of many new Web-based stores and thus a big competitive market for online shopping. Web mining techniques have realized

[2] http://www.inventoryops.com/dictionary.htm

personalized shopping environment, improved direct marketing and advertisement, and enabled companies to understand customers' access patterns for inventory and stock management.

Web mining techniques improve inventory management in (at least) two ways: first, they assist a business in the discovery of vendors or third-party manufacturers; second, those techniques help customers find products that would most likely be suitable to them.

Amazon.com, launched in 1995, is one of pioneer Web sites providing online shopping. Amazon.com earned reputation and popularity for its world-class customer service and its first-rated inventory management. As one of the largest B2C Web sites on earth, Amazon.com needed to put enough effort to both satisfy several million customers as well as efficiently manage inventory.

Within a few years Amazon.com began adopting advanced techniques to manage and plan material resource availability. These techniques enabled Amazon.com to decrease costs and increase the choice of goods for customers, greatly increasing Amazon's revenue. Amazon.com utilizes Web mining techniques to predict customers' demand and thus manage the stocks of different products. Furthermore, taking advantages of Web usage mining techniques and user profiles gathered by a well-designed Web interface, Amazon.com provides personalized interface through personalized recommendations as well as coupons (Gold Box ™).

3.6 Information Technology

The management, construction, and processing of information- and computing-related resources is generally grouped under the business function of information technology. For this, several opportunities exist for businesses to utilize Web mining techniques, such as for reducing the potential for unnecessary duplication in application development.

Developer Duplication Reduction

Many software businesses, both large and small, maintain one or more internal application development units. Thus, at any given time, there may be hundreds, if not thousands, of projects being developed, deployed, and maintained concurrently. Due to overlapping business processes (i.e. human resources and fiscal administration) and multiple project development groups, duplication of source code often occurs (Rajapakse and Jarzabek, 2005) and (Kapser and Godfrey, 2003). Given the non-trivial cost of application development, mitigating such duplication is critical. Source code consistency is also an issue, e.g. to prevent a case where only one of two duplicate segments is updated to address a bug and/or feature addition.

Turnkey solutions for source code duplication are already available, but they suffer from two major problems:

- They are not able to address code which is functionally similar, but syntactically different.
- They only detect duplication after it has already occurred.

The goal of a full-featured duplication detection system will be to address existing and potential duplication – the latter of which is currently unavailable. However, Web mining methods may offer a solution.

Many businesses maintain intranets containing corporate policy information, best practices manuals, contact information, and project details – the last of which is of particular interest here. Assuming project information is kept current, it is possible to use Web mining techniques to identify functionality that is potentially duplicative, oftentimes syntactically different functions may be described using similar language.

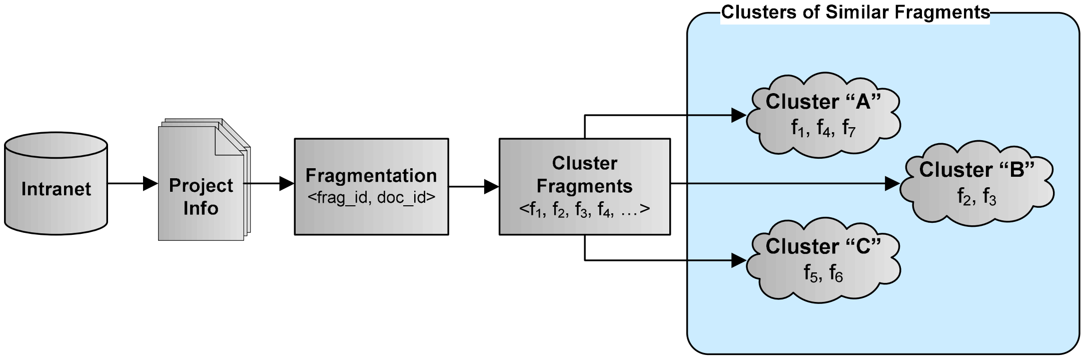


Figure 7: Duplication candidate process overview

Figure 7 gives an overview of a possible approach for identifying potential duplication among multiple projects. First, the project Web pages and documents must be extracted from the intranet. Next, each document is split into fragments using common separators (periods, commas, bullet points, new lines, etc). These fragments form the most basic element of comparison – the smallest entity capable of expressing a single thought. Using clustering techniques, these fragments can then be grouped into collections of similar fragments. When two or more fragments are part of the same collection, but come from different projects, potential duplication has been identified. These fragments may then be red-flagged and brought to the attention of affected project managers.

3.7 Customer Service

The purpose of customer service is to increase the value that buyers receive from their purchases and from the processes leading up to the purchase[3]. What is often different from marketing-related uses of Web mining, however, is that customer service is interested in increasing customer satisfaction by offering correct answers to questions, which means a company's domain knowledge must be leveraged in order to provide a helpful solution.

[3] http://en.wikipedia.org/wiki/Customer_service

Expert Driven Recommendations for Customer Assistance

Most recommender systems used in business today are product-focused, where recommendations made to a customer are typically a function of his/her interests in products (based on his/her browsing history) and that of other similar customers. However, in many cases, recommendations must be made without knowledge about a customer's preferences, like is customer service call centres. In such cases, call centre employees leverage their domain knowledge in order to help align customer inquiries with appropriate answers. Here, a customer may be wrong, which is often observed when domain experts are asked questions by non-experts.

Many businesses must maintain large customer service call centres, especially in retail-based operations, in order to address this need. However, advances in Web-based recommender systems may enable to improve call center capacity by offering expert-based recommendations online (Delong et al, 2005).

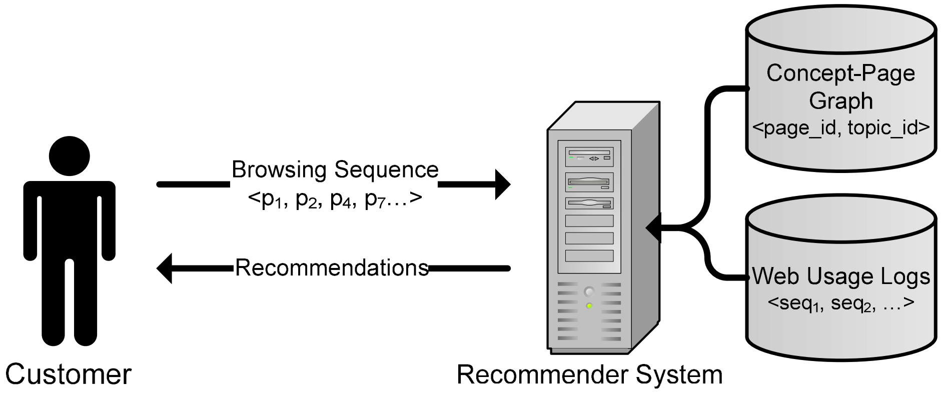


Figure 8: Overview of expert-driven customer assistance recommendations

Similar to a customer talking to a call centre assistant, the recommendation system equates customer browsing behaviour as a series of "questions" the customer wants answered or, more generally, expressions of interest in the topic matter of a clicked-on Web page. Given the order and topic matter covered by such sequences of clicks, the recommendation system continuously refines its recommendations, which are not themselves directly a function of customer interest. Rather, they are generated by querying an abstract representation of customer service Website, called a “concept-page graph”. This graph contains a ranked set of topic/Web page combinations, and as the customer clicks through the Website, the system looks for Web pages best capturing the topics that a customer is seeking to know more about. And since their browsing behaviour helps determine the questions they want answered, the eventual recommendations are more likely find the correct answer to their question, rather than a potentially misleading one based on interest alone.

3.8 Business Process Management

The term Business Process Management (or BPM) refers to activities performed by businesses to optimize and adapt their processes. Any business transaction can be modeled as a sequence of processes that are designed to perform some specific tasks by adding value to the whole transaction. There can be many bottleneck processes in a business transaction and these bottlenecks can severely slow down the whole transaction. It is essential to determine these bottlenecks in a process so that we can redesign the business process model and thereby improve capacity utilization, throughput rate and process time. Business Process Management itself comprises of managing design, execution and monitoring of a process. In the following example, we illustrate how Web mining techniques can be applied to event logs from various processes to develop better models for different processes

Business Process Mining

Business process mining, also called workflow mining, reveals how existing processes work and, thus providing considerable ROI (PMR). Business process mining is the task of extracting useful information from business event logs collected by Workflow Management Systems such as IBM's WebSphere and SAP R/3.

ProM, EMiT and Thumb are some examples of business process mining tools. An example process mining framework is shown in Figure 9.

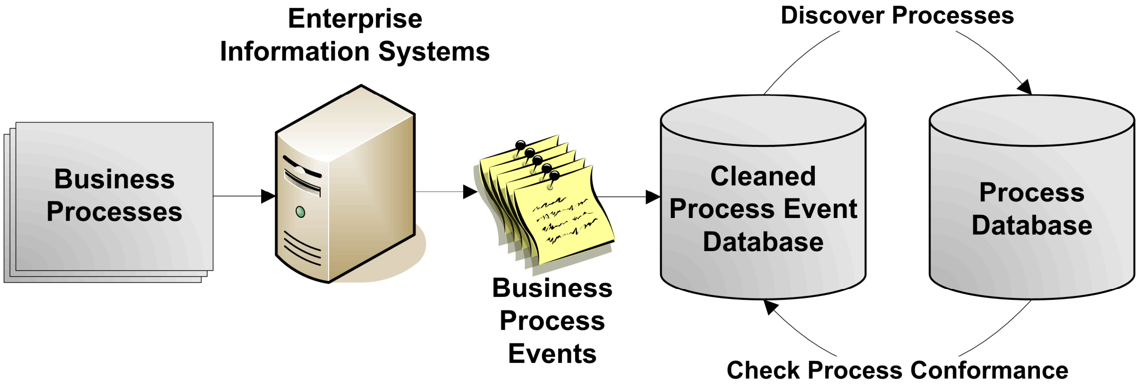

Enterprise Information Systems
Discover Processes
Business Processes
Cleaned Process Event Database
Process Database
Business Process Events
Check Process Conformance

Figure 9: An example process mining framework

Business transaction logs obtained from Enterprise Information systems are transformed into XML format. These event logs are then cleaned and time-ordering of the processes is inferred. Using these, business process models are built and business rules are constructed. Finally, these process models are converted to a Petri-Net for analysis. The process model discovered can also be checked for conformance with previously discovered models. We can also use an anomaly detection system to identify deviation in business process behaviour.

There are many applications like health care management, where business process mining can be used. By digging information from business event logs, for example e-mail traffic, we can discover how people work and interact with each other in an organization. We can see what kinds of patterns exist in workflow processes and answer questions like do people hand-over their tasks to others, do they sub-contract, do they work together or do they work on similar tasks. It thus helps in determining the process, data, organizational and social structure.

Sometimes, the information contained in event logs is incomplete or noisy or fine-grained or specific to an application which makes pre-processing a bit more difficult and challenging. However, by leveraging business process mining properly, we can re-engineer the business process by reducing work-in-progress, adding additional resources to increase the capacity or eliminating or improving the efficiency of bottleneck processes, thereby boosting the performance of businesses.

Table 1. Summary of how Web mining techniques are applicable to different business functions

Function	Application	Technique
Marketing	Product Recommendation	Association Rules
Marketing	Product Trends	Time series data mining
Sales Management	Product sales	Multi-stage supervised learning
Fiscal management	Fraud detection	Link mining
Information Technology	Developer Duplication Reduction	Clustering, Text mining
Customer Service	Expert Driven Recommendations	Association Rules, Text mining, Link Analysis
Shipping and Inventory	Inventory Management	Clustering, Association Rules, Forecasting
Business Process Management	Process Mining	Clustering, Association Rules
Human Resources	HR Call Centers	Sequence similarities, Clustering, Association rules

4 Gaps in Existing Technology

Though Web mining techniques can be extremely useful to businesses, there are gaps which must often be bridged (or completely dealt with) in order to properly leverage Web mining's potential. In this section, we discuss a few such important gaps and how these can be addressed.

4.1 Lack of Data Preparation for Web Mining

To properly apply Web mining in a production setting (e.g. recommending products to customers), data stored in archival systems must be linked back to online applications. As such, there must be processes in place to clean, transform, and move large segments of data back into a setting, where these can be accessed by Web mining applications quickly and continuously. This often means removing extraneous fields and converting textual identifiers (names, products, etc) into numerical identifiers to make the processing of large amounts of transactional data quick. For instance, segmenting data into one-month intervals can cut down on expended computing resources and to ensure that relevant trends are identified by Web mining techniques, provided there is sufficient transactional activity. Additionally, databases for these kinds of intermediate calculations to reduce repeat computations have to be developed. Web mining is often computationally expensive, thus efforts to maximize efficiency are important.

4.2 Under-utilization of Domain Knowledge Repositories

Businesses have long made use of domain knowledge repositories to store information about business processes, policies, and projects, and if they are utilized in a Web mining setting, it becomes ever-more paramount to manage it. For instance, corporate intranets provide a wealth of information that is useful in expert-oriented recommendations (e.g. customer service) and duplication reduction, but the repository itself must be up-to-date and properly maintained from time to time. One of the best ways to ensure an intranet's "freshness" is to maintain it with a content management system (CMS) allowing non-professionals to update the Website and distributing the responsibility to internal stakeholders.

4.3 Under-utilization of Web log data

Most companies keep track of Web browsing behaviour of employees by collecting Web logs mostly for security purposes. However, as seen from previous successful applications of Web mining techniques on such kinds of data, companies could utilize this information to better serve their employees. For example, one of the key issues that is usually dealt by human resources department is to keep employees motivated and retain them. A common approach is to offer perks and bonuses in various forms to satisfy the employee. However, most policies are ‘corporate-centric’ and are not geared towards ‘employee-centric’. With the advance of Web mining techniques, it is now possible to understand employees’ interests in a better way. Two kinds of techniques can be employed. First, is to mine the behaviour of employees in company policy and benefits Website, in order to understand what employees are looking for. For example, employees browsing retirement benefits related Website, could be offered a better retirement package. Other examples include, tuition waiver for employees looking on pursuing professional development course, or a travel package deal to an employee who has shown interest in travelling. A different dimension is to use trend analysis to see what’s new and popular in market, such as a new mp3 player, and offer perks in form of such as products. Ofcourse, a key issue in such kind of techniques is privacy. Privacy preserving data mining is a currently a hot area of research. Also, it has been studied and shown from examples such as Amazon that people are willing to compromise a certain level of privacy to gain the benefits offered.

5 Looking Ahead: The Future of Web Mining in Business

We believe that the future of Web mining is entwined with the emerging needs of businesses, and the development of techniques fuelled by the recognition of gaps or areas of improvement in existing techniques. This section examines what is on the horizon for Web mining, the nascent areas currently under research, and how they can help in a business computing setting.

5.1 Microformats

It is very important to not only to present the right content on a Web site, but also in the right format. For example, a first step in formatting for the Web was the use of <HTML> to give the browser’s ability to parse and present text in a more readable and presentable format. However, researchers soon developed formats with higher semantics and presentability, e.g. XML and CSS, for efficient processing of content and extracting useful information. XML is used to store data in formats such that automatic processing can be done to extract meaningful information (not just for presenting it in a Website). Today, the trend is moving more towards “micro-formats” which capture the best of XML and CSS. Microformats are design principles for formats and not another new language. They provide a way of thinking about data, which will provide humans a better understanding of the data. They are currently widely used in Websites such as blogs. With such new structured data, there arises need for NLP and Web content mining techniques such as data extraction, information integration, knowledge synthesis, template detection and page-segmentation. This leads to the suggestion for the corporate businesses to decide on right kind of format to best utilize the data for processing, analysis and presentation.

5.2 Mining and Incorporating Sentiments

Even though automated conceptual discovery from text is still relatively new, difficult, and imperfect, accurately connecting that knowledge to sentiment information – how someone feels about something – is even harder. Natural language processing techniques, melded with Web mining, hold great promise in this area. To understand how someone feels about a particular product, brand, or initiative, and to project that level of understanding across all customers would give the business a more accurate representation of what customers think to date. Applied to the Web, one could think of an application that collects such topic/sentiment information from the Internet, and returns that information to a business. Accomplishing this would open up many marketing possibilities.

5.3 eCRM to pCRM

Traditionally, brick-and-mortar stores have been organized in a product-oriented manner, with aisles for various product categories. However, success of online e-CRM initiatives in the on-line world in building customer loyalty is not hidden from CRM practitioners in the physical world, which we refer to as p-CRM for clarity in this chapter. Additionally, the significance of physical stores has motivated a number of online businesses to open physical stores to serve “real people” (Earle 2005). Many businesses have also moved from running their online and physical stores separately to integrating both, in order to better serve their customers (Stuart 2000). Carp (2001) points out that although the online presence of a business does affect its physical division of its business, people still find entertainment value in shopping in malls and other physical stores. Finally, people prefer to get a feel of products before purchase, and hence prefer to go out to shop instead of shopping online. From these observations, it is evident that physical stores will continue to be the preferred means of conducting consumer commerce for quite some time. However, margins will be under pressure as they must adopt to compete with online stores. These observations led us to posit the following in our previous study (Mane et al 2005b):

> “Given that detailed knowledge of an individual customer’s habits can provide insight into his/her preferences and psychology, which can be used to develop a much higher level of trust in a customer-vendor relationship, the time is ripe for revisiting p-CRM to see what lessons learned from e-CRM are applicable.”

Till recently, a significant roadblock in achieving this vision has been the ability to collect and analyze detailed customer data in the physical world, as Underhill’s seminal study (Underhill 1999) showed, both from cost and customer sensitivity perspectives. With advancements in pervasive computing technologies such as mobile Internet access, third-generation wireless communication, RFIDs, handheld devices and Bluetooth; there has been a significant increase in the ability to collect detailed customer data. This raises the possibility of bringing e-CRM style real-time, personalized, customer relationship functions to the physical world. For a more detailed study on this , refer to our previous work (Mane et al 2005b)

5.4 Other directions

We have mentioned some of the key issues that should be noted by businesses as they proceed to adopt Web mining techniques to improve the business intelligence. However, as claimed our earlier, this by no means is an exhaustive list. There are various other issues that need to be addressed from technical perspective of Web mining to determine necessary framework to make these techniques applicable in businesses. For example, there are host of open areas of research in Web mining techniques, such as extraction of structured data from unstructured data or ranking of Web pages integrating the semantic relationship within documents and sentiments of the users. Also, businesses have to focus on the kind of data that needs to be collected for use of Web usage mining techniques to be effective. Designing the framework of Website plays a crucial role in deciding what kind of data can be collected. For example, one viewpoint is that pages with Flash, though are very attractive, are more of broadcast nature and do not explicitly collect information about customer's interests. However, recent technologies such as AJAX, involve increased customer interaction, which not only allows corporations to collect data, but also gives the customer a 'sense of control' leading to better satisfaction.

6 Conclusion

This chapter examines how technology, such as Web mining, can aid businesses in gaining an extra information and intelligence. We provide an introduction to Web mining and the various techniques associated with it. We briefly update the reader with state-of-art research in this area. Later, we show how these class of techniques can be effectively used to aid various business functions. We provide example applications to illustrate the use of Web mining techniques to aid in certain areas of business functions. These examples provide the evidence of success and the potential of Web mining for business intelligence. Finally, we point out the gaps in existing technologies and certain future directions that should be of interest to the business community at large. In doing so, we also note that we have intentionally left out the technical directions of future work, with the view of the target audience.

7 Acknowledgements

This work was supported in part by AHPCRC contract number DAAD19-01-2-0014, by NSF Grant ISS-0308264 and by ARDA grant F30602-03-C-0243. This work does not necessarily reflect the position or policy of government and no official endorsement should be inferred. We would like to thank the Data Mining Research Group at the University of Minnesota for providing valuable feedback.

References

[1] Adomavicius, G. and Tuzhilin, A., Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowledge and Data Engineering, Vol. 17, 2005.

[2] Apte, C., Liu, B., Pednault, E.P.D., and Smyth, P., Business Applications of Data Mining, Communications of the ACM, 45(8), August 2002.

[3] Arsham, H., Time-Critical Decision Making for Business Administration, http://home.ubalt.edu/ntsbarsh/stat-data/Forecast.htm, 2006.

[4] Bolton, R. and Hand, D., Statistical fraud detection: A review, Statistical Science, Vol 17(3), pp. 235-255, 2002.

[5] Bose,A; Beemanapalli,K; Srivastava, J; Sahar,S, "Incorporating Concept Hierarchies into Usage Based Recommendations", WEBKDD 2006, August 20, 2006,Philadelphia, Pennsylvania, USA

[6] Box, G. E. and Jenkins, G. M., Time Series Analysis: Forecasting and Control, 3rd. Prentice Hall PTR, 1994.

[7] Carp, J., Clicks vs. Bricks: Internet Sales Affect Retail Properties., In Houston Business Journal, Feb.16, 2001.

[8] Cooley, R., Mobasher, B. and Srivastava, J., Data preparation for mining World Wide Web browsing patterns. Knowledge and Information Systems, 1(1), pp. 5–32, 1999.

[9] Cooley, R., Mobasher, B. and Srivastava, J., Web mining: information and pattern discovery on the World Wide Web. 9th IEEE ICTAI 1997.

[10] Cortes, C., Pergibon, D. and Volinsky, C., Communities of Interest. Lecture Notes in Computer Science, 2189, 105-114, 2001.

[11] DeLong, C., Desikan, P., and Srivastava, J., USER (User Sensitive Expert Recommendation): What Non-Experts NEED to Know. Proceedings of WebKDD, Chicago, Illinois, 2005.

[12] Desikan, P., Srivastava, J., Kumar, V. and Tan, P. N., Hyperlink Analysis: Techniques and Applications. Technical Report 2002-0152, Army High Performance Computing and Research Center, 2002.

[13] Earle, S., From Clicks to Bricks...Online Retailers Coming Back Down to Earth. Feature story. http://www.specialtyretail.net/issues/december00/feature_bricks.htm, 2005.

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

[15] Getoor, L., Link Mining: A New Data Mining Challenge. SIGKDD Explorations, 4(2), 2003.

[16] Kapser, C. and Godfrey, M. W., Toward a Taxonomy of Clones in Source Code: A Case Study. International Workshop on Evolution of Large-scale Industrial Software Applications, Amsterdam, The Netherlands, 2003.

[17] Keogh, E., Data Mining in Time Series Databases Tutorial. Proceedings of the IEEE Int. Conference on Data Mining, 2004.

[18] Kleinberg, J.M., Authoritative Sources in Hyperlinked Environment. 9th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 668-667, 1998.

[19] Kosala, R. and Blockeel, H., Web mining research: A survey. SIGKDD Explorations, 2(1), pp. 1 - 15, 2000.

[20] Liu, B. and Chang, K.C.C., Editorial: Special Issue on Web Content Mining. SIGKDD Explorations special issue on Web Content Mining, Dec, 2004.

[21] Mane, S., Desikan, P., and Srivastava, J., From Clicks to Bricks: CRM Lessons from E-commerce. Technical report 05-033, Department of Computer Science, University of Minnesota, Minneapolis, USA, 2005b.

[22] Mane, S., Vayghan, J., Srivastava, J., Yu, P., and Adomavicius, G., Data Mining Techniques for Automated Evaluation of Sales Opportunities: A Case Study. International Workshop on Customer Relationship Management: Data Mining Meets Marketing, 2005.

[23] Mobasher, B., Cooley, R., and Srivastava, J., Automatic Personalization Based on Web Usage Mining. Communications of ACM, August 2000.

[24] Mobasher, B., Web Usage Mining and Personalization. Practical Handbook of Internet Computing, ed. M.P. Singh, CRC Press, 2005.

[25] Mullins, C. S., Gaining Knowledge Through Process Mining. http://www.enterpriseleadership.org/read/mullins2

[26] Norris, B, What is Marketing?, http://www.briannorris.com/whatismarketing.html

[27] Page, L., Brin, S., Motwani, R. and Winograd, T., The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies, January 1998.

[28] Perkowitz, M. and Etzioni, O., Adaptive Web sites: an AI challenge. IJCAI 1997.

[29] Ph.H.B.F Franses, Modeling New Product Sales; An Application of Co-Integration Analysis. International Journal of Research in Marketing, 1994.

[30] Phua C, Lee V, Smith K, and Gayler R., A Comprehensive Survey of Data Mining-based Fraud Detection Research. Artificial Intelligence Review, 2005.

[31] Pirolli, P., Pitkow,J.E., Distribution of Surfer's Path Through the World Wide Web: Empirical Characterization. World Wide Web 1:1-17, 1999.

[32] PMR, Process Mining Research, http://www.processmining.org

[33] Rajapakse, D. C. and Jarzabek, S., An Investigation of Cloning in Web Applications. Fifth International Conference on Web Engineering, Sydney, Australia, July 27-29, 2005.

[34] Sarukkai, R.R., Link Prediction and Path Analysis using Markov Chains. Proc. of the 9th World Wide Web Conference, 1999.

[35] Senator, T. et al., The financial crimes enforcement network AI system (FAIS) – Identifying potential money laundering from reports of large cash transactions. AI Magazine, 16, 21-39, 1995.

[36] Srivastava, J. and Mobasher, B., Panel discussion on "Web Mining: Hype or Reality?" ICTAI 1997.

[37] Srivastava, J., Desikan, P. and Kumar, V., Web Mining - Concepts, Applications and Research Directions (Book Chapter). Data Mining: Next Generation Challenges and Future Directions, MIT/AAAI 2004.

[38] Stuart, A., Clicks and Bricks. In CIO Magazine, March 15th, 2000.

[39] Underhill, P., Why We Buy: The Science of Shopping. Simon and Schuster, New York, 1999.

[40] Van der Aalst, W.M.P. , Weijters, A.J.M.M and Maruster, L., Workflow Mining: Discovering Process Models from Event Logs, IEEE Transactions on Knowledge and Data Engineering (TKDE), volume 16(9), pages 1128-1142, 2004.

[41] Van der Aalst, W.M.P. , van Dongen, B.F. , Herbst, J. , Maruster, L. , Schimm, G. and Weijters, A.J.M.M. Workflow Mining: a Survey of Issues and Approaches, Data and Knowledge Engineering, 47(2):237-267, 2003.

[42] Van Dongen, B.F., de Medeiros, A.K.A., Verbeek, H.M.W. , Weijters A.J.M.M. and van der Aalst, W.M.P. The ProM framework: A new era in process mining tool support. 26th International Conference on Applications and Theory of Petri Nets, 2005.

[43] Zhu, J., Hong, J. and Hughes, J.G., Using Markov Chains for Link Prediction in Adaptive Web Sites. Proc. of ACM SIGWEB Hypertext, 2002.