ABSTRACT:
Many users of the internet are aware that each time they connect to an online shopping server, they leave behind a “footprint” in the site’s server logs. The information contained in this footprint is innocuous, but it can be ‘mined’. Data mining is the ‘important technique of determining valid, original, useful and eventually understandable patterns in data’.

Keywords: Data Mining, Electronic Commerce, Web mining, Business Intelligence.

[1] INTRODUCTION
This paper is to increase understanding of the determinants for adoption of e-commerce in data mining. Also it shows that there is a positive relationship between e-commerce and data mining. In the first section we discuss importance of both in today’s business world. In the second section we discuss about the current scenario of online shopping. Third section explains the architecture of data mining. Fourth section discusses the data collection techniques. Further we describe the data sources in online shopping & data mining implementations in ecommerce.


[2.1] IMPORTANCE OF E-COMMERCE IN TODAY'S BUSINESS WORLD
The leading position for business today is Electronic Commerce (E-commerce). Most people assume E-commerce means online purchasing. But online purchasing is only a subset of the E-commerce. In addition, E-commerce includes business-to-business networks that make procuring easier for large business organizations. E-commerce generally described as a method of buying and selling things and services online. The main building blocks of E-commerce are the Internet and the World Wide Web, email, fax, and telephonic orders.

Main ingredient of e-commerce is information processing. The e-commerce effects in all areas of business, from client service to new goods design. It assists new types of methods for attaining and collaborating with clients – online advertising and marketing, online-order taking
and online client service etc. It can also decrease costs in handling orders. Also E-commerce provides the new types of information and innovative goods such as interactive games, e-books, and information-on-demand that can be very useful for content providers and consumers.

[2.2] IMPORTANCE OF DATA MINING IN TODAY’S BUSINESS WORLD

Data and Information or Knowledge has an eloquent aspect on human activities. Data mining is the process by evaluating large amount of data from various panoramas and encapsulating it into fruitful information. Due to the importance of obtaining knowledge/information from the abundant data archives, data mining has transformed into an imperative element in various fields of human life.

Data Mining is by large used in several areas like understanding consumer research marketing, good analysis, demand and supply analysis, and e-commerce, investment trend in stocks & real estates, telecommunications and so on. Data Mining has great significance in today’s highly competing business environment. A new idea of Business Intelligence data mining has occurred, which is mostly used by prominent corporate houses to stay ahead of their competitors. Business Intelligence (BI) can help in giving latest data. It is used for market research, consumer behavior, economical trends, industry research, geographical information analysis, competition analysis and so on. Business Intelligence Data Mining assists in decision-making. Famous data mining software programs are Connexor Machines, Free Text Software Technologies, Megaputer Text Analyst, SAS Text Miner, LexiQuest, WordStat, Lextek Profiling Engine.

[3] ONLINE SHOPPING IN INDIA HAS HIT A TIPPING POINT

Online research conducted by TNS says online shopping in India saw 128% growth in interest from the consumers in the year 2011 to 2012, compared to only 40% in 2010 to 2011. Google India's report says 90% of online shoppers are planning to buy more goods online.

There are many companies which support the feature of e-transaction which in turn attracts the client. Clients just need to register in/on the respective website and the door for e-transaction gets opened. Ipsos, one of the global market research company conducted online interviews of 12,000 people in September 2012 across 24 countries, with adults aged 18-64. The US data is based on a sample size of 500. The survey revealed that around 40% people avoid doing online transactions.

The main reason is that the online store values user privacy and confidentiality. We have surveyed many web sites for this. Here are the examples-

*Pepperfry.com* is one of the Best Online Shopping Website in India. They need personal information like mobile number, gender for sign up an account.

*HushBabies.com* is India's one-stop shopping source for parents and parents-to-be. They also need personal information like mobile number, date of birth for sign up an account.
The above examples shows that these kind of websites needs personal details while registration like mobile number, SSN(Social Security Number), credit card number etc and whenever you checkout, it is likely that the company is a scam or they want these kind of data only for demographic purpose.

[4] AN ARCHITECTURE FOR DATA MINING

In an e-commerce setting, it is very much expected that vendors, clients and application service providers (ASP) (usually the middlemen) have different data mining needs. Merchants would be keen in data mining meant for market basket analysis to know client segments, whereas end clients are keen to know updates on regular offerings. The role of the application service providers is then to be the common assembly ground for merchants and clients. Krishnaswamy et al recommend a dispersed data mining architecture that capacitates a data mining to be done in such a naturally distributed environment [1]. In sequence to pillar the strong operation of the system it acquires certain qualities such as heterogeneity, costing infrastructure availability, presence of optimization engine, security and extensibility. Heterogeneity means that the process can mine data from disparate and scattered sites. The recommended system is designed to hold user requirements w.r.t various scattered computing standards (including the client-server and mobile agent based models). The costing framework refers a task that demands advanced computational assets and/or expeditious feedback time should cost the users more on a comparative degree of costs. Moreover, the system must be able to use effectively the scattered data mining process to benefit the users with the best feedback time possible.

Preserving security means that in some cases, the user may be searching highly sensitive data that should not leave the owner's site. In such cases, the authors give the privilege to make use of the mobile-agent model where the mining algorithm and the appropriate guidelines are inserted to the data site and at the end of the process the mobile agent is demolished on the site itself. The system is capable of extension to provide for a wide range of mining algorithms [1]. The authors furnish an opportunity wherein the user can sign up their algorithms with the ASP for use in their particular distributed data mining jobs. After gathering bulk data, comes the role of data mining. In case of online shopping, data mining helps in understanding buying patterns of a client, their choices, and preferences. The data given by a particular client gets stored in web logs, cookies etc. Such type of data is mined for corrective decision making.

[5] DATA COLLECTION

[5.1] ENABLING DATA COLLECTION IN E-COMMERCE

It may be observed that there are different ways of acquiring data appropriate to e-commerce DM. Web server log files, web server plug-ins(instrumentation), TCP/IP packet sniffing, application server instrumentation are the basic ways of gathering raw data. Other ways consists of transactions that the user do, marketing programs, demographic, call centers and ERP
Systems. It is quite general to exhaust about 80% of any DM exercise in e-commerce in data cleaning. This is largely in part to the heavy reliance on the web logs that are generated by the HTTP protocol. This protocol being stateless, it becomes very problematic to take out client’s purchasing behavior-related information together with the good technicalities Ansari et al. describe an architecture for supporting the integration of DM and e-commerce [11]. The architecture focused on the requirement of data collection at the application server layer and not the web server, in order to support tagging of data and metadata that is imperative to the discovery process. They also explain the data transformation bridges required from the transaction processing systems and client event streams to the data warehouse.

[5.2] ANALYZING WEB TRANSACTIONS

Once the data are aggregated, data analysis could be done. Data analysis could be done with the help of session level, client, good and abstract attributes.

Session level analysis maintains the details related to the number of page views, unique pages, time spent, average, fast versus slow connection per session etc. Also, session level analysis could highlight on whether users did registration, if so, when, did they see the privacy statement feature; did they use inquiry features, etc. The user level analysis could highlight whether the user is an initial or repeat or recent visitor/purchaser; whether the users are readers, browsers, heavy spenders, original referrers etc. [11]

In web-usage analysis, Markov chains have been used to develop presumptions between page views and to decrease the system latencies.

Hu&Ceren gives a new technique called on-line analytical mining for web data [9]. This technique consists of many phases like data capture phase, web house construction phase, and pattern discovery and evaluation phase. This technique is beneficial in finding the most effective clients, the difference between patrons and non-patrons, labeling of website menus that drag many of the visitors, menus of website that are session killers, menus of the web site that influence to the most of the buyers, finding the difficult path of clients that traverses to a buy or otherwise etc. The web house is directly connected to the data warehouse.

[6] DATA SOURCES IN ON LINE SHOPPING

[6.1] SERVER RECORDS

On line shopping servers build many network protocols which have the feasible to be mined. The common log format (CLF) gives data about physical connections by this syntax: ‘host ident authored date request status byte’. The CLF gives records about logical connections, as well as software and hardware profile of the consumer. Generally, the data of the creator will be in the form of an IP (Internet Protocol) number. By the help of reverse DNS (Domain Name Server), it is feasible to find the full domain name which can then be analyzed. However it must be find out that not all domain names can be confirmed.
[6.2] COOKIE RECORDS

HTTP (Hyper Text Transfer Protocol) is a state less protocol i.e. each transaction work independently. The stateless nature of HTTP is well suited to transaction processing. For example, the online shopping trolleys store your specific purchases, even if you close that website and re-visit the same after several days or weeks later. Cache facility in the web server may store all the requests that are generated by the visitors. Cookies are normally used to save information like the contents of a shopping trolley, or the pages accessed when a client last connected to a website. The same request may be there then it must be responded by the last re-visited page through cookies records. This will save the response time. Cookie records contain data in the form ‘name expiry_date, path domain and security_level’, which can be tailored depending on the applied domain.

[6.3] CLIENT INFORMATION

The data that is gathered using on-screen forms to empower shielded credit authorization is a rich originator of supplementary data about the client. This process is usually prefaced by the identification process of the user, which can either be done at the first usage of the on-line service, or at the first good buy. The type of appealed data depends on the type of the on-line shopping business and the outlook of the usage of the data.

[6.4] MISCELLANEOUS SOURCES

[6.4.1] DEMOGRAPHIC DATA

It is general practice to gather demographic data from third party collaborators. These data has usually been collected over decades and provides valuable information for strategic judgments. Database marketing has become a very profitable field, which essay such data, often conformed for definite requirements.

[6.4.2] INTERNET SOURCES

A contemporary data source is the Internet itself, which has been taken advantage by various parties. Useful material about users is cataloged and recommended for acceptance to concerned clients, either through intelligent agents (and equivalents) or a cookie, an example of a company which provides such information is DoubleClick.

DoubleClick Ad Exchange is a real-time digital ad marketplace that connects publishers with agencies, ad networks and third-party technology providers.
[7] DATA MINING IMPLEMENTATIONS IN E-COMMERCE

[7.1] CLIENT PROFILING

It may be seen that clients drive the earnings of any organization. Accessing new clients, gratifying and holding current clients, and forecasting client demeanor will enhance the availability of goods and services and hence the profits. Thus the end goal of any data mining exercise in e-commerce is to improve processes that contribute to delivering value to the end buyer. Consider an online store like http://www.dell.com where the client can configure a PC of his/her choice, place an order for the same, track its movement, as well as pay for the good and services. With the technology behind such a web site, dell has the opportunity to make the retail experience exceptional. At the most basic level, the information available in web log files can predict what prospective clients are seeking from a site.

Companies like Dell provide their clients access to details about all of the systems and configurations they have purchased so they can incorporate the information into their capacity planning infrastructure integration. Back-end technology systems for the website include sophisticated data mining tools that take care of knowledge representation of client profiles and predictive modeling of scenarios of client interactions. For example, once a client has purchased a certain number of servers, they are likely to need additional routers, switches, load balancers, back up devices etc. rule mining based systems could be used to propose such alternatives to the clients.

[7.2] RECOMMENDATION SYSTEMS

Systems have also been developed to keep the clients accordingly updated of big events which are of relevance to them. The article by Jeng & Drissi [2] discusses an exceptional framework called PENS that has the ability to not only notify clients of events, but also to predict events and events classes that are likely to be activated by clients. The event notification system in PENS has the following components: event manager, event channel manager, registries and proxy manager. The event-prediction system is based on association rule mining and clustering algorithms. The PENS system is used to strongly help an e-commerce service provider to predict the requirements of good categories better. Data mining has also even exercised in identifying how clients may react to promotional offers made by a credit card e-commerce company [3]. Applications including fuzzy computing and interval computing are used to develop if-then-else rules.

[7.3] WEB PERSONALIZATION

Mobasher presents a broad survey of the complete process based on web usage mining [4]. In this context, the author discusses a host of web usage mining projects vital for this process, including the preparing and combining of data from many sources, and common pattern finding techniques such as clustering, association rule-mining and sequential pattern discovery, performed
on web usage data, can be influenced completely as a main part of a web personalization system. The author observes that the log data possessed inevitably by the web and application servers represent the exquisite exploration behavior of visitors.

Depending on the goals of the analysis, e-commerce data need to be converted and gathered at different levels of consideration. E-commerce data are also further divided as usage, content, structure and user data. Usage data consists of details of user sessions and page views. The content data in a site are the combination of objects and relationships that are conducted to the user. For the most part, the data consists of collection of texture material and pictures. The data sources used to transfer or produce data include static HTML/XML pages, pictures, video clippings, sound files, dynamically produced page segments from scripts or other applications, and accumulation of records from the operational database(s). Site content data also include semantic or structural metadata contained within the site or individual pages, such as detailed keywords, document attributes, semantic tags, or HTTP variables. Structure data depicts the designer’s perspective of the content organization within the site. This organization is held via the inter-page architecture of the content depicted in the compact of HTML or XML tags within a page. Structure data for a site are grabbed by an auto-generated site map include additional user profile data. Such data may include demographic or other identifying information on registered users, user ratings on various objects such as pages, goods, or movies, past purchase or visit histories of users, as well as other explicit or implicit representations of users’ interests.

[7.4] DM & BUYER BEHAVIOR IN E-COMMERCE

For a successful e-commerce site, suppressing user-recognized dormancy is the second highly essential characteristic after good site-exploration characteristic. The most efficacious towards suppressing user-recognized dormancy has been the uprooting of route navigation impressions from former user’s approach experiences to determine upcoming user navigation demeanor and to prefect the required assets. However, this proposal is appropriate for only non e-commerce sites where there is no buying demeanor. Vallamkaondu & Gruenwald illustrate a way to forecast user demeanor in e-commerce sites [5]. The heart of their proposal involves uprooting knowledge from correlative data of purchase and route navigation impressions of past users (obtainable from web server logs) to predict the buying and routing demeanor of upcoming users.

Web sites are generally to portray an organization’s image, to advertise and market goods and to administer client help. The eminence of a web site influence and mirrors straightly the eminence of the organization in the e-market. Spiliopoulou & Pohle propose a mechanism to enhance the eminence of web sites, based on the exploitation of navigation pattern discovery [6]. In particular, the authors present theory, in which success is modeled on the basis of the navigation behavior of the sites’ users, they then exploit web usage miner (WUM), a navigation pattern discovery miner, to study how the success of a site is reflected in the users’ behavior. With WUM the authors measure the success of a site’s components and obtain concrete indications of how the site should be improved.

In the context of web mining, clustering could be used to cluster similar click-streams to
determine learning behaviors in the case of e-learning or general site access behaviors in e-commerce. Most of the algorithms presented in the literature to deal with clustering web sessions treat sessions as sets of visited pages within a time period and do not consider the sequence of the click stream visitation. This has significant consequences when comparing similarities between web sessions. Wang & Zaiane propose an algorithm based on sequence alignment to measure similarities between web sessions where sessions are chronologically ordered sequences of page accesses [7].

[7.5] DM AND MULTIMEDIA E-COMMERCE

Operations in fundamental multimedia enumerations are extremely correlative, as in e-malls auctioning multimedia content based goods. It is crucial in such circumstances to measure asset’s requirements needed for delivering of catalog chapters. Hollfelder et al nominate a mechanism to figure out delivering asset requirements in correlative multimedia catalogs [10]. The divination depends on the outcomes of mining the fundamental mall action log file that stores data about former user’s activity like ordering and purchasing demeanor.

[8] CONCLUSION

The statement “E-commerce: the killer Domain for Data Mining” is optimal because many of the components required for blooming data mining are easily accomplished: data records are abundant, electronic collection impart factual data; perception can fluently be swung into force. To really take leverage of this discipline, however, data mining must be incorporated into the e-commerce systems with the pertinent data changeover branched from the transaction processing system to the data warehouse and vice-versa. Such integration can dramatically dwindle the data preparation time, known to take about 80% of the time to conclude an evaluation. An integrated key can also cater users with a consistent user interface and seamless approach to metadata.
REFERENCES


