ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

High- through functions and classification of Data Mining Inter-specific with Web Mining

A.K Soniyapriyadharishni* and Dr.P.B.Ramesh babu**

***Research Scholar, Department of Bioinformatics, Bharath University, India.**
****Professor and HOD, Department of Bioinformatics, Bharath University, India.**

ABSTRACT

From its very beginning, the potential of extracting valuable knowledge from the Data and the Web has been quite evident. We well know that, Data mining is a technique where we get the extract of raw data and useful information, Whereas, Web mining is the application of data mining techniques to extract knowledge from Web content, structure, and usage. Interest in Data mining and Web mining has grown rapidly in its short existence, both in the research and practitioner communities. This paper provides a brief overview of the accomplishments that "How Data mining is Inter-specific with Web mining" fields both in terms of technologies and applications - and outlines key functions, classification and research directions.
Keywords: Content Mining, Data Mining, Deployment, Usage Mining, Web Mining.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

Whereas, Web mining **-** is the application of data mining techniques to discover patterns from the Web. Thus to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. The attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences.

What is data mining?

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related - also known as "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).

Stage 1: Exploration.

This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and - in case of data sets with large numbers of variables ("fields") - performing some preliminary feature selection operations to bring the number of variables to a manageable range (depending on the statistical methods which are being considered). Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage.

Stage 2: Model building and validation.

This stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable

results across samples). This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models," that is, applying different models to the same data set and then comparing their performance to choose the best. These techniques - which are often considered the core of predictive data mining - include: Bagging (Voting, Averaging), Boosting, Stacking (Stacked Generalizations), and Meta-Learning.

Stage 3: Deployment.

That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome. The concept of *Data Mining* is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques specifically designed to address the issues relevant to business *Data Mining* (e.g., Classification Trees), but Data Mining is still based on the conceptual principles of statistics including the traditional Exploratory Data Analysis (EDA) and modeling and it shares with them both some components of its general approaches and specific techniques.

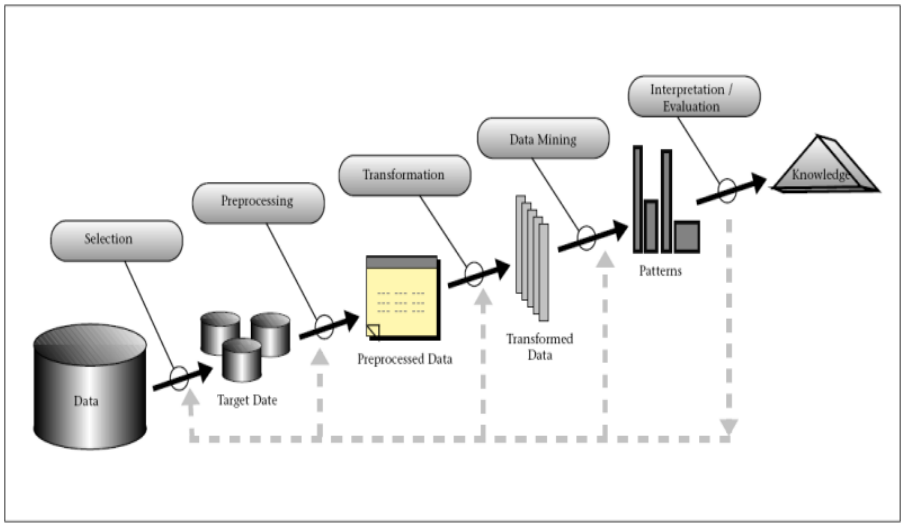


Fig 1. The data mining process

Decisive features in Data Mining

Bagging (Voting, Averaging)

The concept of bagging (voting for classification, averaging for regression-type problems with continuous dependent variables of interest) applies to the area of predictive data mining, to combine the predicted classifications (prediction) from multiple models, or from the same type of model for different learning data. It is also used to address the inherent instability of results when applying complex models to relatively small data sets.

Suppose your data mining task is to build a model for predictive classification, and the dataset from which to train the model (learning data set, which contains observed classifications) is relatively small. You could repeatedly sub-sample (with replacement) from the dataset, and apply, for example, a tree classifier to the successive samples. In practice, very different trees will often be grown for the different samples, illustrating the instability of models often evident with small data sets. One method of deriving a single prediction is to use all trees found in the different samples, and to apply some simple voting: The final classification is the one most often predicted by the different trees.

Boosting

The concept of boosting applies to the area of predictive data mining, to generate multiple models or classifiers, and to derive weights to combine the predictions from those models into a single prediction or predicted classification.

A simple algorithm for boosting works like this: Start by applying some method (e.g., a tree classifier such as C&RT or CHAID) to the learning data, where each observation is assigned an equal weight. Compute the predicted classifications, and apply weights to the observations in the learning sample that are inversely proportional to the accuracy of the classification. In other words, assign greater weight to those observations that were difficult to classify (where the misclassification rate was high), and lower weights to those that were easy to classify (where the misclassification rate was low). In the context of C&RT for example, different misclassification costs (for the different classes) can be applied, inversely proportional to the accuracy of prediction in each class. Then apply the classifier again to the weighted data (or with different misclassification costs), and continue with the next iteration (application of the analysis method for classification to the re-weighted data).
Boosting will generate a sequence of classifiers, where each consecutive classifier in the sequence is an "expert" in classifying observations that were not well classified by those preceding it. During deployment (for prediction or classification of new cases), the predictions from the different classifiers can then be combined (e.g., via voting, or some weighted voting procedure) to derive a single best prediction or classification.

Data Preparation (in Data Mining)

Data preparation and cleaning is an often neglected but extremely important step in the data mining process. The old saying "garbage-in-garbage-out" is particularly applicable to the typical data mining projects where large data sets collected via some automatic methods (e.g., via the Web) serve as the input into the analyses. Often, the method by which the data where gathered was not tightly controlled, and so the data may contain out-of-range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), and the like. Analyzing data that has not been carefully screened for such problems can produce highly misleading results, in particular in predictive data mining.

ISSN: 0975-8585

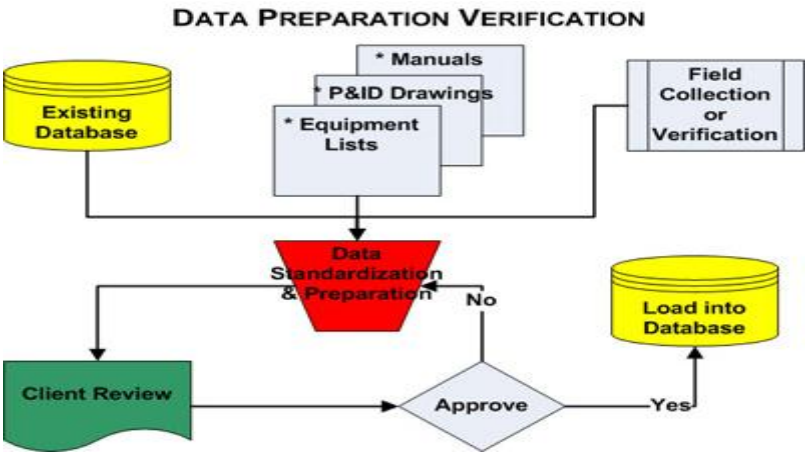


Fig 2. Data preparation in data mining

Data Reduction (for Data Mining)

The term Data Reduction in the context of data mining is usually applied to projects where the goal is to aggregate or amalgamate the information contained in large datasets into manageable (smaller) information nuggets. Data reduction methods can include simple tabulation, aggregation (computing descriptive statistics) or more sophisticated techniques like clustering, principal components analysis, etc.

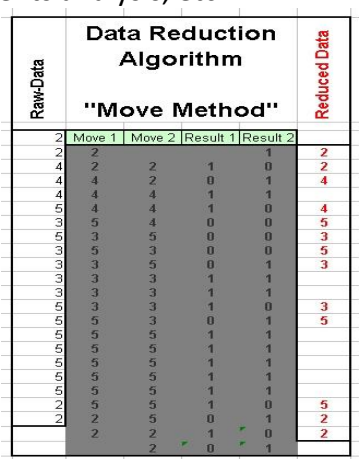


Fig 3. Data Reduction algorithm in Data mining

Deployment

The concept of deployment in predictive data mining refers to the application of a model for prediction or classification to new data. After a satisfactory model or set of models has been identified (trained) for a particular application, we usually want to deploy those models so that predictions or predicted classifications can quickly be obtained for new data. For example, a credit card company may want to deploy a trained model or set of models (e.g., neural networks, meta-learner) to quickly identify transactions which have a high probability of being fraudulent.

Drill-Down Analysis

The concept of drill-down analysis applies to the area of data mining, to denote the interactive exploration of data, in particular of large databases. The process of drill-down analyses begins by considering some simple break-downs of the data by a few variables of interest (e.g., Gender, geographic region, etc.). Various statistics, tables, histograms, and other graphical summaries can be computed for each group. Next, we may want to "drill-down" to expose and further analyze the data "underneath" one of the categorizations, for example, we might want to further review the data for males from the mid-west. Again, various statistical and graphical summaries can be computed for those cases only, which might suggest further break-downs by other variables (e.g., income, age, etc.). At the lowest ("bottom") level are the raw data: For example, you may want to review the addresses of male customers from one region, for a certain income group, etc., and to offer to those customers some particular services of particular utility to that group.

Feature Selection

One of the preliminary stage in predictive data mining, when the data set includes more variables than could be included (or would be efficient to include) in the actual model building phase (or even in initial exploratory operations), is to select predictors from a large list of candidates. For example, when data are collected via automated (computerized) methods, it is not uncommon that measurements are recorded for thousands or hundreds of thousands (or more) of predictors. The standard analytic methods for predictive data mining, such as neural network analyses, classification and regression trees, generalized linear models, or general linear models become impractical when the number of predictors exceed more than a few hundred variables.
Feature selection selects a subset of predictors from a large list of candidate predictors without assuming that the relationships between the predictors and the dependent or outcome variables of interest are linear, or even monotone. Therefore, this is used as a pre-processor for predictive data mining, to select manageable sets of predictors that are likely related to the dependent (outcome) variables of interest, for further analyses with any of the other methods for regression and classification.

Machine Learning

Machine learning, computational learning theory, and similar terms are often used in the context of Data Mining, to denote the application of generic model-fitting or classification algorithms for predictive data mining. Unlike traditional statistical data analysis, which is usually concerned with the estimation of population parameters by statistical inference, the emphasis in data mining (and machine learning) is usually on the accuracy of prediction (predicted classification), regardless of whether or not the "models" or techniques that are used to generate the prediction is interpretable or open to simple explanation. Good examples of this type of technique often applied to predictive data mining are neural networks or meta-learning techniques such as boosting, etc. These methods usually involve the fitting of very complex "generic" models, which are not related to any reasoning or theoretical understanding of underlying causal processes; instead, these

techniques can be shown to generate accurate predictions or classification in cross validation samples.

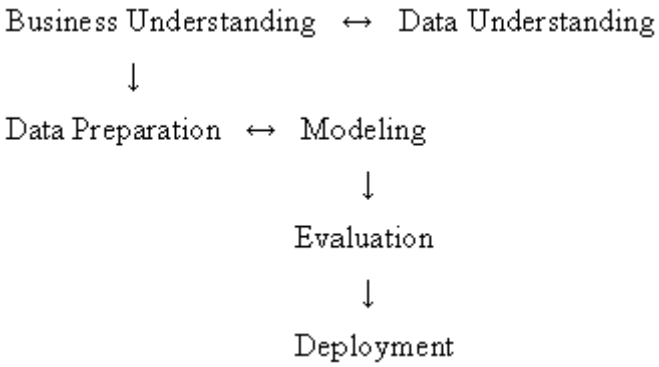
Meta-Learning

The concept of meta-learning applies to the area of predictive data mining, to combine the predictions from multiple models. It is particularly useful when the types of models included in the project are very different. In this context, this procedure is also referred to as Stacking. Suppose your data mining project includes tree classifiers, such as C&RT and CHAID, linear discriminant analysis (e.g., see GDA), and Neural Networks. Each computes predicted classifications for a cross validation sample, from which overall goodness-of-fit statistics (e.g., misclassification rates) can be computed. Experience has shown that combining the predictions from multiple methods often yields more accurate predictions than can be derived from any one method. The predictions from different classifiers can be used as input into a meta-learner, which will attempt to combine the predictions to create a final best predicted classification. So, for example, the predicted classifications from the tree classifiers, linear model, and the neural network classifier(s) can be used as input variables into a neural network meta-classifier, which will attempt to "learn" from the data how to combine the predictions from the different models to yield maximum classification accuracy.

Models for Data Mining

In the business environment, complex data mining projects may require the coordinate efforts of various experts, stakeholders, or departments throughout an entire organization. In the data mining literature, various "general frameworks" have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements.

One such model, CRISP (Cross-Industry Standard Process for data mining) was proposed in the mid-1990s by a European consortium of companies to serve as a non-proprietary standard process model for data mining. This general approach postulates the following (perhaps not particularly controversial) general sequence of steps for data mining projects:



Another approach - the Six Sigma methodology - is a well-structured, data-driven methodology for eliminating defects, waste, or quality control problems of all kinds in manufacturing, service delivery, management, and other business activities. This model has recently become very popular (due to its successful implementations) in various American industries, and it appears to gain favor worldwide. It postulated a sequence of, so-called, DMAIC steps –

$$\text{Define} \rightarrow \text{Measure} \rightarrow \text{Analyze} \rightarrow \text{Improve} \rightarrow \text{Control}$$

that grew up from the manufacturing, quality improvement, and process control traditions and is particularly well suited to production environments (including "production of services," i.e., service industries).

Another framework of this kind (actually somewhat similar to Six Sigma) is the approach proposed by SAS Institute called SEMMA –

$$\text{Sample} \rightarrow \text{Explore} \rightarrow \text{Modify} \rightarrow \text{Model} \rightarrow \text{Assess}$$

This is focusing more on the technical activities typically involved in a data mining project.

All of these models are concerned with the process of how to integrate data mining methodology into an organization, how to "convert data into information," how to involve important stake-holders, and how to disseminate the information in a form that can easily be converted by stake-holders into resources for strategic decision making.

What is Web Mining?

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. 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, 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. The second definition has become more acceptable, as is evident from the approach adopted. 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. We can see the taxonomy of web mining in Fig. 4

Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery,

extracting association patterns, clustering of web documents and classification of Web Pages.

Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to Web content mining has been limited.

Web Structure Mining

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting related pages. Web Structure Mining is the process of discovering structure information from the Web. This can be further divided into two kinds based on the kind of structure information used.

Hyperlinks: A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink. There has been a significant body of work on hyperlink analysis, of which 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.

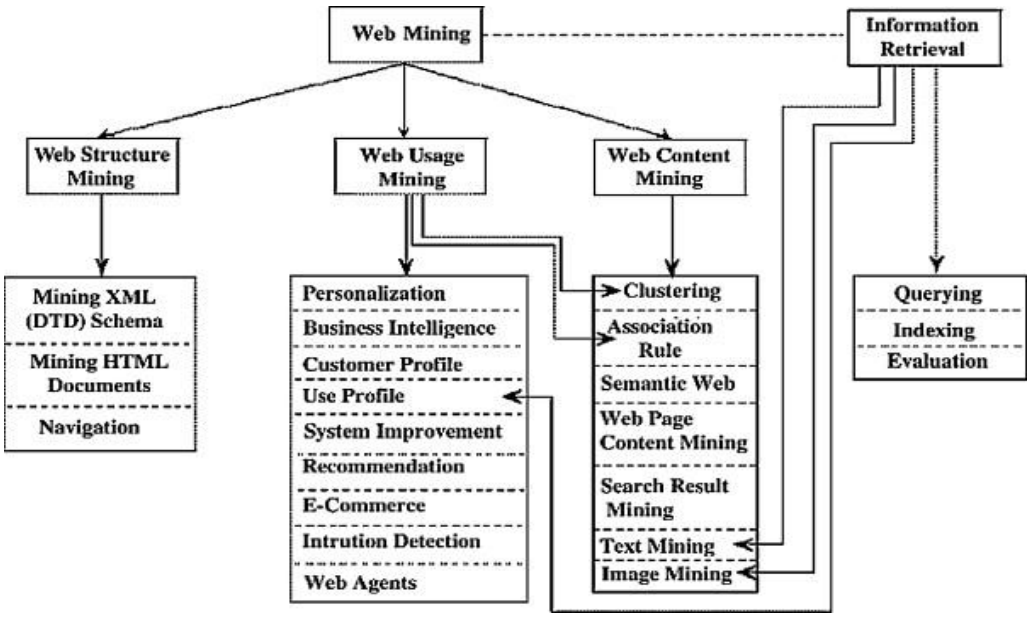


Fig 4 . Taxonomy of Web Mining

ISSN: 0975-8585

Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to under- stand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

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

Application Server Data: Commercial application servers such as Web logic, Story Server 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 specially defined 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.

Decisive features in Web mining

In this section we briefly describe the key new features of the Web mining database

Ranking metrics - for page quality and relevance

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

PageRank

PageRank(Fig.5) is a metric for ranking hypertext documents based on their quality developed this metric for the popular search engine Google/Yahoo. The key idea is that a page has a high rank if it is pointed to by many highly ranked pages. So, the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively till the rank of all the pages is determined. The rank of a page p can thus be written as:

$$PR(p) = d/n + (1-d) \sum_{(G.P)\sum G} (PR(q)/outdegree(q)$$

ISSN: 0975-8585

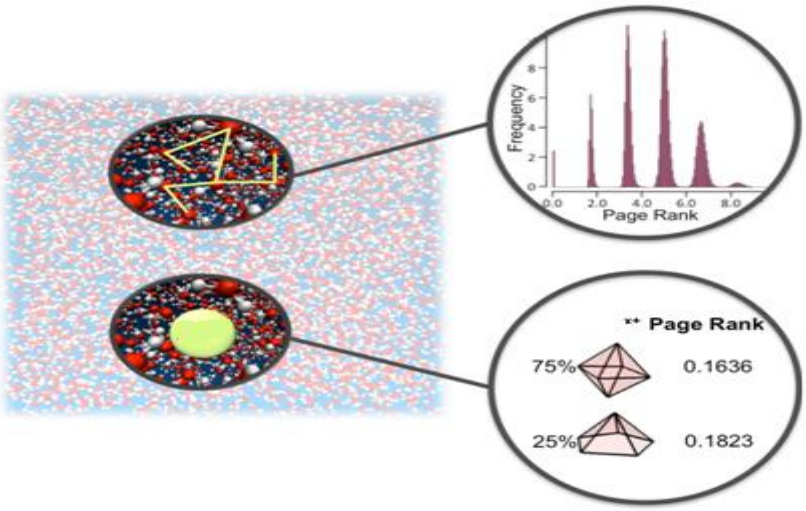


Fig 5. Page ranking with frequency

Hubs and Authorities:

Hubs and Authorities can be viewed as 'fans' and 'centers' in a bipartite core of a Web graph, where the nodes on the left represent the hubs and the nodes on the right represent the authorities. The hub and authority scores computed for each Web page indicate the extent to which the Web page serves as a "hub" pointing to good "authority" pages or as an "authority" (Fig.6) on a topic pointed to by good hubs. The hub and authority scores are computed for a set of pages related to a topic using an iterative procedure called HITS.

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 is 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. 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 Page Rank and Hubs and Authorities approaches, Topic Sensitive Page Rank and Web page Reputations.

Robot Detection and Filtering - Separating human and nonhuman Web

Behavior Web robots are software programs that automatically traverses 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 model. First of all, e-commerce retailers are particularly concerned about the unauthorized deployment of robots for gathering business intelligence at their Web sites. In addition, Web robots tend to consume considerable network bandwidth at the expense of other users. Sessions due to

ISSN: 0975-8585

Web robots also make it difficult to perform click-stream analysis effectively on the Web data. Conventional techniques for detecting Web robots are often based on identifying the IP address and user agent of the Web clients. While these techniques are applicable to many well-known robots, they are not sufficient to detect camouflaged and previously unknown robots. An approach that uses the navigational patterns in click-stream data to determine if it is due to a robot. Experimental results have shown that highly accurate classification models can be built using this approach. Furthermore, these models are able to discover many camouflaged and previously unidentified robots.

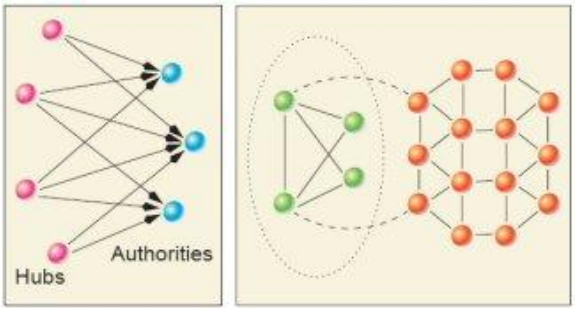


Fig 6. Hubs and Authorities

Information scent - Applying foraging theory to browsing behavior

Information scent is a concept that uses the snippets and information presented around the links in a page as a "scent" to evaluate the quality of content of the page it points to, and the cost of accessing such a page. 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.

User profiles - Understanding how users behave

The Web has taken user profiling to completely new levels. For example, in a 'brick and-mortar' store, data collection happens only at the checkout counter, usually called the 'point-of-sale'. This provides information only about the final outcome of a complex human decision making process, with no direct information about the process itself. In an on-line

store, the complete click-stream is recorded, which provides a detailed record of every single action taken by the user, providing a much more detailed insight into the decision making process. Adding such behavioral information to other kinds of information about users, e.g. demographic, psychographic, etc., allows a comprehensive user profile to be built, which can be used for many different applications.

Interestingness measures - When multiple sources provide

Conflicting evidence one of the significant impacts of publishing on the Web has been the close interaction now possible between authors and their readers. In the pre-Web era, a reader's level of interest in published material had to be inferred from indirect measures such as buying/borrowing, library checkout/renewal, opinion surveys, and in rare cases feedback on the content. For material published on the Web it is possible to track the precise click-stream of a reader to observe the exact path taken through on-line published material.

We can measure exact times spent on each page, the specific link taken to arrive at a page and to leave it, etc. Much more accurate inferences about readers' interest in content can be drawn from these observations. Mining the user click-stream for user behavior, and using it to adapt the 'look-and-feel' of a site to a reader's needs.

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

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 repository of publications. Some of the prominent digital libraries are SCI, ACM portal, CiteSeer and DBLP. Fig.7 represents the process of online bibiliometrics.

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. (Fig.8)

ISSN: 0975-8585

Fig 7. Online bibiliometrics in web mining

Analyzing the web log data with visualization tools has evoked a lot of interest in the research community. Chi et al 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”. The most 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.

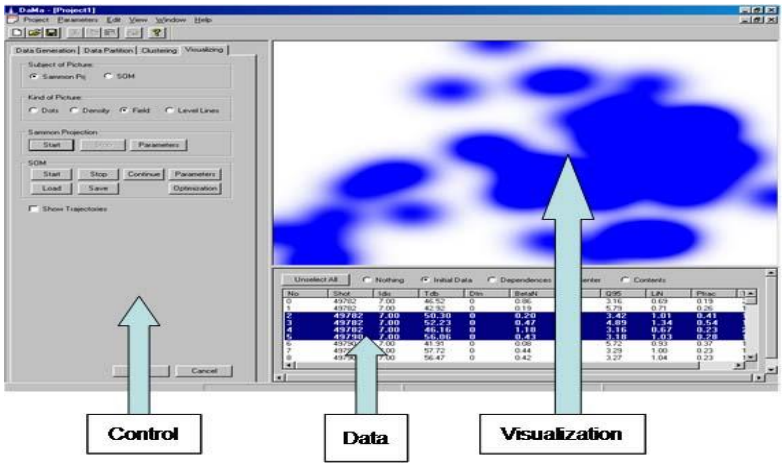


Fig 8. Visualization of World Wide Web

All of these models are concerned with the process of how to integrate web mining methodology into an organization, how to "convert data into information," and the information into the web mining concept. And thus, how to involve important stake-holders, and how to disseminate the information in a form that can easily be converted by stake-holders into resources for strategic decision making.

ISSN: 0975-8585

Data mining inter-specific with Web mining

In data mining the computerization and automated data gathering has resulted in extremely large data repositories. E.g., Walmart: 2000 stores, 20 M transactions/day, where the raw data patterns the knowledge. The scalability issues and desire for more automation makes more traditional techniques less effective.eg. Statistical Methods, Relational Query Systems, OLAP. Whereas, in web mining the attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences since here the potential of extracting valuable knowledge from the web, 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. The major differences between data mining and web mining are given below in the table. (Table 1).

- Data mining has been classified into three ways: The initial exploration, Model building or pattern identification with validation/verification, and Deployment, whereas web mining also classified into three ways such as: Web content, Web structure and Web Usage.

- Clustering in data mining has been done in two simple different ways that is Partitional Clustering where a division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset and Hierarchical clustering where a set of nested clusters organized as a hierarchical tree. Followed by data mining clustering in web mining has also done in two different simple ways “Supervised” technique where categories are defined and documents are assigned to one or more existing categories the “definition” of a category is usually in the form of a term vector that is produced during a “training” phase Training is performed through the use of documents that have already been classified (often by hand) as belonging to a category and “Unsupervised” technique where the documents are divided into groups based on a similarity metric, no pre-defined notion of what the groups should be, Most common similarity metric is the dot product between two document vectors.

- Analysis in data mining can be greatly studied in statistics and neural network fields Examples: Predicting sales amount of a new product based on advertising expenses and time series prediction of stock market indices. Perhaps analysis in web mining can be studied in combination of clustering and classification As new documents are added to a collection, an attempt is made to assign each document to an existing topic (category), where, the collection is also checked for the emergence of new topics and similarly the drift in the topics are also identified.

- The hierarchy concept in mining the data starts from the data sample to explore and modify the model and validate further to access the knowledge, whereas the hierarchy in mining the web starts with categories hence organizations of categories leads to documentation and then further leads to category dimensions. Thus, Organization of categories; e.g. Flat, Tree, or Network, and Category Dimensions; e.g. Subject, Location, Time, Alphabetical, Numerical.

- At the final stage of mining the content of data mining is revealed with deviation and detection by discovering most significant changes in data from previously measured or normative data. Usually categorized separately from other data mining tasks and the deviations are often infrequent. Thus, modifications of classification, clustering and time series analysis can be used as means to achieve the goal and to outlier detection in statistics. Where the relevance in web mining, can be measured with respect to any of the following criteria i.e. Documentation, Query based, User Based, Role/Task Based. Thus, these kind of similar ways make data mining inter-specific with web mining.

Table 1. Differences between Data Mining and Web Mining

Contents	Data mining	Web mining
Definition	Data Mining is an analytic process designed to explore data.	Web mining is the application of data mining techniques to discover patterns from the Web.
Technique	Raw data->Patters->Knowledge	Knowledge data-> Web->Collection of files-> Web server
Classification	Exploration, Modelling, Deployment	Content, Structure, Usage
Clustering	Partitional/Hierarchical	Supervised/Unsupervised
Topic analysis	Prediction / Marketing	Identification / Tracking
Concept hierarchy	Sample-> Explore-> Modify-> Model-> Access	Organization of categories-> Documentation-> Category dimension
Content Relevance	Deviation and Detection	Documenting, Query based, User based and Task based

CONCLUSION

Thus, from the above discussion we got a clear idea how data mining is inter specific with Web mining technology. Thus as we said, Data Mining is an analytic process designed to explore data and from the outcome of the excitement of web mining. Where, web applications have been developed at a much faster rate in the industry than research in Web related technologies. Many of these were based on the use of data mining concepts even though the organizations that developed these applications, and invented the corresponding technologies. 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 and decisive features and contributions made by the field. Our hope is that this overview provides a starting point for fruitful discussion.

REFERENCES

[1] Broderet al. In the Proc. 9th WWW Conference 2000.

[2] J Borges, M Levene. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), New York City, 1998.

[3] S Brin, L Page, In the 7th International World Wide Web Conference, Brisbane, Australia, 1998.

[4] Broadvision1-to-1 portal, http://www.bvportal.com/.

[5] D Clark. Shopbots become agents for business change, IEEE Computer, 18-21. R. Cooley, Web Usage Mining: Discovery and Usage of Interesting Patterns from Web Data, Ph.D. Thesis, University of Minnesota, Computer Science & Engineering, 2000.

[6] Colet, DSStar, 2002.

[7] R Cooley, B Mobasher, . Srivastava. In Proceedings of the 9th International Conference on Tools With Artificial Intelligence (ICTAI ’97), Newport Beach, CA, 1997.

[8] M Pazzani, L Bguyen, S Mantik. In Proceedings of the International Conference on Tools with AI, 1995.

[9] M Pazzani, J Muramatsu, D Billsus. In Proceedings of AAAI/IAAI Symposium, 1996.

[10] L Page, S Brin, R Motwaniand, T Winograd. Stanford Digital Library Technologies, 1999-0120, January 1998.

[11] M Perkowitz, O Etzioni. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 1999.

[12] B Prasetyo et al. Pacific-Asia Conference on Knowledge Discovery and Data Mining 2002, Taipei, Taiwan.

[13] Pandey J, Srivastava S, Shekhar. SIAM Workshop on Web Mining, 2001.

[14] Padmanabhan A, Tuzhilin. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York, NY,1998.

[15] M Spiliopoulou. Proceedings of the Symposium on Principles of Knowledge Discovery in Databases (PKDD).