Research Journal of Recent Sciences ISSN 2277 - 2502
Vol. **1(ISC-2011),** 375-387 **(2012)** *Res.J.Recent.Sci.*

Review Paper

Study and Evaluation of user's behavior in e-commerce Using Data Mining

Belsare Satish and Patil Sunil
Department of Computer Science, SCMIPS, Indore, MP, INDIA

Available online at: www.isca.in
(Received 15th October 2011, revised 17th January 2012, accepted 25th January 2012)

Abstract

Data mining has matured as a field of basic and applied research in computer science. The objective of this dissertation is to evaluate, propose and improve the use of some of the recent approaches, architectures and Web mining techniques (collecting personal information from customers) are the means of utilizing data mining methods to induce and extract useful information from Web information and service where data mining has been applied in the fields of e-commerce and e-business (that means User's behavior). 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 a significant consequence when comparing similarities between web sessions. Wang and Zaiane propose an algorithm based on sequence alignment to measure similarities between web sessions where sessions are chronologically ordered sequences of page accesses.

Keywords: User behavior, e-commerce, web mining, clustering, data mining.

Introduction

User Behavior: In addition to understanding the needs of the customers, Company also need to understand what motivates them to purchase, and how can influence the buying process to ensure that the products or services are on the shopping list.

Understanding the customers will help, to develop and distribute the product, as well as getting the right price point and developing successful promotional activities.

The psychology of the buying process has been widely studied and no matter what size company business, knowledge of this process can help company become more successful.

Both businesses and consumers exhibit patterns of buying behavior. The business model is less open to debate as the business customers will almost certainly have some formalized process of buying in place. The company task is to understand the process and match the marketing activities to the different stages of the process. This means that the customer will receive the right kind of contact at the right time.

To provide insights into areas such as indicators of customer defection, price sensitivity, segmentation, and customer needs analysis, to name a few.

Our R and D work in this area has exposed many commonly held marketing beliefs as mere mythologies with no valid scientific basis. We employ this unique knowledge to answer questions have such as: What are the leading indicators of customer defection in our industry? What indicates the 'health' of our brand? And what does this mean for us? How can predictive modeling help me determine the outcome of our marketing actions? What are customer's reservation prices in our industry? How can we use this to predict the impact of price changes? What are our customer's needs and how do factors such as age or gender impact? Which brands do we compete most closely with?

Business Buying Behavior: A typical business customer will go through the following steps when buying: Identifying a need or problem: This may be highlighted by press coverage or advertising they have seen in the trade press. Developing product specification**:** The customer will use whatever sources they can find to help them specify what they need. They will pay particular attention to press releases, exhibitions, advertising, editorial comment, industry seminars and relevant direct mail. Search for products and suppliers: This is the time when the business customer is particularly open to visits from your sales force and trade directory entries. Exhibitions and technical information leaflets are also invaluable sources. This is the time when pricing information begins to be seriously considered.

Evaluation of products and suppliers: This is a good time to provide your potential customer with demonstration products, visits to existing customers, plant visits or third party testimonials. You may also need to look at special pricing packages or stocking incentives.

Ready to place an order: This is the time for personal contact.

Evaluation of product and supplier performance: The more major the buying decision, the more reassurance your customer needs. Review meetings and helpline support provide reassurance, as does good after sales support and continued exposure to advertising and press coverage - justifying the purchase decision.

Follow on purchase: The first purchase should not be seen as the end of the process, but the beginning of a long-term business relationship.

Consumer Buying Behavior: There are many models of consumer buying behavior, but the steps below are fairly common to most of them.

The customer identifies a need: This is often initiated by PR coverage, including word of mouth. The customer may have seen a friend or celebrity using a product or service, or awareness may have been sparked off by advertising.

Looking for information: At this stage the customer wants to know more and is actively seeking information. Advertising and PR are still important but product demonstrations, packaging and product displays play a role. This is the time to deploy your sales personnel, and customers find videos and brochures are useful. Word of mouth is still very important.

Checking out alternative products and suppliers: The customer is now trying to choose between products, or firm up on the purchase decision. This is a place for promoting product guarantees and warranties, and maximizing packaging and product displays. Sales personnel can greatly influence the customer at this stage and sales promotion offers become of interest. Independent sources of information are still of interest, including product test reviews.

Purchase decision: This is the time to 'tip the balance'. Sales promotion offers come into their own, and if appropriate, sales force incentives need to ensure that your sales personnel are incentives to close the deal.

Using the product: Expensive purchases can lead to what is known as cognitive dissonance - a fear that the customer has not made the right decision. Your job is to reassure the customer by offering good customer care, simple instruction manuals and loyalty schemes. They should still be exposed to testimonial advertising to reassure them that they have made the right decision.

Marketing does not stop at understanding the buying processes of the customer however, company need to understand their buying patterns and the market in which they operate.

User Behavior and Data Mining: Most marketers understand the value of collecting customer data, but also realize the challenges of leveraging this knowledge to create intelligent, proactive pathways back to the customer. Data mining technologies and techniques for recognizing and tracking patterns within data – helps businesses sift through layers of seemingly unrelated data for meaningful relationships, where they can anticipate, rather than simply react to, customer needs. In this accessible introduction, Kurt Thearling provides a business and technological overview of data mining and outlines how, along with sound business processes and complementary technologies, data mining can enforce and redefine customer relationships.

Data Mining and Customer Relationships: The way in which companies interact with their customers has changed dramatically over the past few years. A customer's continuing business is no longer guaranteed. As a result, companies have found that they need to understand their customers better, and to quickly respond to their wants and needs. In addition, the time frame in which these responses need to be made has been shrinking. It is no longer possible to wait until the signs of customer dissatisfaction are obvious before action must be taken. To succeed, companies must be proactive and anticipate what a customer desires.

It is now a cliché that in the days of the corner market, shopkeepers had no trouble understanding their customers and responding quickly to their needs. The shopkeepers would simply keep track of all of their customers in their heads, and would know what to do when a customer walked into the store. But today's shopkeepers face a much more complex situation. More customers, more products, more competitors, and less time to react means that understanding the customers is now much harder to do. A number of forces are working together to increase the complexity of customer relationships:

Compressed marketing cycle times: The attention span of a customer has decreased dramatically and loyalty is a thing of the past. A successful company needs to reinforce the value it provides to its customers on a continuous basis. In addition, the time between a new desire and when customer must meet that desire is also shrinking. If company doesn't react quickly enough, the customer will find someone who will.

Increased marketing costs: Everything costs more. Printing, postage, special offers (and if company doesn't provide the special offer, the competitors will).

Streams of new product offerings: Customers want things that meet their exact needs, not things that sort-of fit. This means that the number of products and the number of ways they are offered have risen significantly.

Niche competitors: The best customers also look good to the competitors. They will focus on small, profitable segments of the market and try to keep the best for themselves.

Successful companies need to react to each and every one of these demands in a timely fashion. The market will not wait for company response, and customers that have today could vanish tomorrow. Interacting with the customers is also not as simple as it has been in the past. Customers and prospective customers want to interact on their terms, meaning that company need to look at multiple criteria when evaluating how to proceed. Company will need to automate: The Right Offer, To the Right Person, At the Right Time, Through the Right Channel.

The right offer means managing multiple interactions with the customers, prioritizing what the offers will be while making sure that irrelevant offers are minimized. The right person means that not all customers are cut from the same cloth. The company interactions with them need to move toward highly segmented marketing campaigns that target individual wants and needs. The right time is a result of the fact that interactions with customers now happen on a continuous basis. This is significantly different from the past, when quarterly mailings were cutting-edge marketing. Finally, the right channel means that company can interact with the customers in a variety of ways (direct mail, email, telemarketing, etc.). The company needs to make sure that are choosing the most effective medium for a particular interaction.

The purpose of this dissertation is to provide company with a thorough understanding of how a technology like data mining can help solve vexing issues in the interactions with the customers. We describe situations in which a better understanding of the customers can provide tangible benefits and a measurable return on investment. It is important to realize, though, that data mining is just a part of the overall process. Data mining needs to work with other technologies (for example, data warehousing and marketing automation), as well as with established business practices. If company takes nothing else from this dissertation, we hope that the company will appreciate that data mining needs to work as part of a larger business process (and not the other way around).

Data Mining: Data mining, by its simplest definition, automates the detection of relevant patterns in a database. For example, a pattern might indicate that married males with children are twice more likely to drive a particular sports car than married males with no children. If any persons are a marketing manager for an auto manufacturer, this somewhat surprising pattern might be quite valuable.

However, data mining is not magic. For many years, statisticians have manually "mined" databases, looking for statistically significant patterns. Data mining uses well-established statistical and machine learning techniques to build models that predict customer behavior. Today, technology automates the mining process, integrates it with commercial data warehouses, and presents it in a relevant way for business users. The leading data mining products are now more than just modeling engines employing powerful algorithms. Instead, they address the broader business and technical issues, such as their integration into today's complex information technology environments. In the past, the hyperbole surrounding data mining suggested that it would eliminate the need for statistical analysts to build predictive models. However, the value that an analyst provides cannot be automated out of existence. Analysts will still be needed to assess model results and validate the plausibility of the model predictions. Because data mining software lacks the human experience and intuition to recognize the difference between a relevant and an irrelevant correlation, statistical analysts will remain in high demand.

Web Data Mining: Web data mining is one kind of these techniques that efficiently handle the tasks of searching the needed information from the Internet, improving the Web site structure to provide better Internet service quality and discovering the informative knowledge from the Internet for advanced Web applications. Web data mining could be categorized into three types of Web content, Web structure[1] and Web usage mining[2,3,4]. In this study, we focus on Web usage mining: that is, discovering user access pattern knowledge from Web log files, which contain the historic visiting records of users on the website.

Web Content Mining: It is an automatic process that goes beyond keyword extraction. Since the content of a text document presents no machine readable semantic, some approaches have suggested restructuring the document content in a representation that could be exploited by machines. There are two groups of web content mining strategies: Those that directly mine the content of documents and those that improve on the content search of other tools like search engines. Web Structure Mining: Worldwide Web can reveal more information than just the information contained in documents Web Usage Mining: Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of deferent web sites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking.

The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analyses can shed light on better structure and grouping of resource providers. Many web analysis tools existed but they are limited and usually unsatisfactory. We have designed a web log data mining tool, Weblog Miner, and proposed techniques for using data mining and OnLine Analytical Processing (OLAP) on treated and transformed web access files. Applying data mining techniques on access logs unveils interesting access patterns that can be used to restructure sites in a more efficient grouping, pinpoint effective advertising locations, and target specific users for specific selling ads. Customized usage tracking analyzes individual trends. Its purpose is to customize web sites to users. The information displayed the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns. While it is encouraging and exciting to see the various potential applications of web log file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge one can discover from the large raw log data. Current web servers store limited information about the accesses. Some scripts custom-tailored for some sites may store additional information. However, for an effective web usage mining, an important cleaning and data transformation step before analysis may be needed.

Technical issues: Benefits for the companies: Today, the enormous content of the Internet has made it difficult to find relevant information on a subject. Methods helping user navigation and retrieving information have become particularly important. Online shops need to offer personalized products to clients but before being able to do that they have to personalize the web sites to the clients. This is where the data mining techniques in web server logs are coming in. Companies can use the basic data retrieved from the data logs to analyze customer behaviors, evaluate the current usage, if the customers liked or disliked it and so on. To create adaptable web sites to each user, first, the user navigation patterns in the web have to be found and analyzed. Data mining is a method extracting valuable information from the data for statistical purpose.

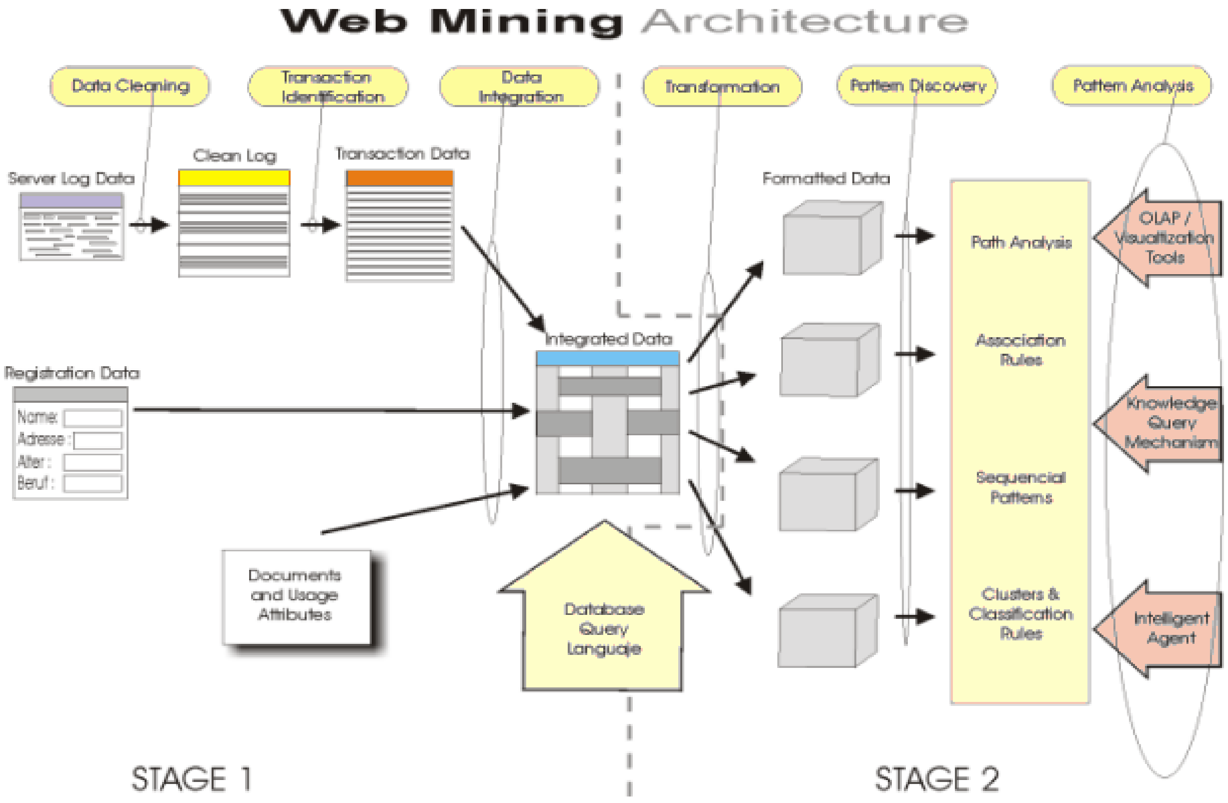


Figure-1
Architecture of Web mining

How does it work? Data mining techniques: One of the most known data mining techniques is **WEBSOM** (web based self-organizing map), which organizes documents into two-dimensional map according to their content rather than by keyword.

Unlocking the ***usage patterns*** of the web users hidden in the log files has therefore been a challenge for several researchers. Fu, Perkowitz and Etzioni[5] have demonstrated that web users can be grouped into meaningful clusters, which help the web designers to provide high-value customized services as the data mining of the web server logs provides them with the information needed to understand users better. These systems can also be used to improve current WebPages.

Mobasher[6,7] created a Web Personalization system that organizes web usage data not the content of the data mentioned above into clusters. The system analyzes the web server logs, it identifies to which user group the current user belongs to and makes suggestions to links that would interest the user. These suggestions are based on the past experience of a particular user group.

LOGSOM (log based self-organizing map) combines the benefits of the both of these systems. It keeps a track and organizes the web pages according to the user navigation behaviors and interest not to the web content[8].

More detailed view how do they actually do it? Data caching algorithms: The general idea is to make the full use of the web log data using data mining applications. Data mining is aimed to discover the standards, structure, content, usage patterns and so on of users and web pages. Intelligent web caching algorithms are the tools to predict the web requests. “These web-caching algorithms are able to adapt their behavior on the basis of the user access patterns, which in step are extracted from the historical access data recorded in the log files by means of data mining techniques.”

The goal of these algorithms is to increase the number of web pages that are retrieved directly from the cache instead requesting them from the server. There are several web caching algorithms, we list three of them:

Frequent patterns In the case of frequent patterns, we extract from the web logs the patterns that follow the form A→B (if A, then B). If A has been requested then B is likely to be requested next.

Decision trees In the case of decision trees, we develop a decision tree in a basis of the historical data in the web logs but on this case concentrating on the time needed until the next request.

Page gather This algorithm uses the data mining clustering to group a collection of web sites that are regularly visited close to each other and distant from other groups. Data mining clustering differs from the traditional clustering in a way that it can place one document in a multiple overlapping clusters.

Which type of data can be collected from where? Everything you do while surfing in the Net can provide useful information, for instance, for web site designers.

Clickstream data is the path that the user creates when steering through the sites and following links. It can be used to evaluate the traffic and popularity of the page

Shopping chart can provide information in e-business where the purchases were made and where the customer left the order unfinished.

Psychographic data would include data on user’s attitudes towards topics, products etc., buying behavior and beliefs.

Access data counts the time between the last and next access to the same URL.

Time data gives information on amount of time a user spends exploring the site, the product or topic he or she is interested in.

All that provides us with the useful information about the users but how far can we go until it becomes a privacy concern? Is it appropriate to record all the user activities in order to find out how users perceive the site?

Social issues: What Do Users Think? As data mining tools and algorithms become more sophisticated and widely available, customer’s privacy concerns are constantly increasing. These concerns are especially high due to opportunity of World Wide Web to easily automatically collect consumer data and add it to databases. With organizations increasingly building comprehensive consumer databases and applying sophisticated data-mining techniques privacy and ethics issues become more pressing.

The user’s think about the following four questions[1,9], Does Internet Data Mining Violate Users’ Privacy? What is the User Persistence of Internet Data Mining? What can marketers do about Internet Data Mining Privacy? What can users do about Internet Data Mining Privacy.

From the results above 4 tips for online marketers can be suggested to improve user trust and data mining techniques’ privacy:

Explain the purpose of data mining: Data mining always has a certain goal. “The survey found that at least some segments of online users lose their negative approach to Internet data mining when they better understand the purpose of it".

Control of information distribution: Users are in majority strongly against the sharing of mined data among different companies.

Provide key trust points to improve e-commerce practice: That implies clear statements of privacy policy. Also privacy software and description or communication about that.

Adjust privacy methods in respect of different customer groups: Customers have different opinions and concerns. e.g. "increased age generally correlates with increased concerns about online privacy. Those are the tips for companies to improve their privacy performance. But are there any tools for users to protect themselves from access without notification to their personal information by online companies.

Why mine e-commerce and click stream data? Improve conversion rate through personalization. Optimize marketing campaigns (banners, email, and other media) that bring visitors to your site by measuring return on investment (ROI). Improve basket size through cross-sells and up-sells. Streamline navigation paths through the site. Avoid content delivery issues (poorly formatted for AOL, too rich for low bandwidth users, redundant or confusing content). Identify customers segments that you can target offline. Experiment quickly. The Web is a laboratory. Understand what works quickly.

Material and Methods

Clustering Algorithm: Clustering analysis is a widely used data mining algorithm for many data management applications. Clustering is a process of partitioning a set of data objects into a number of object clusters, where each data object shares the high similarity with the other objects within the same cluster but is quite dissimilar to objects in other clusters. Different from classification algorithm that assigns a set of data objects with various labels previously defined via a supervised learning process, clustering analysis is to partition data objects objectively based on measuring the mutual similarity between data objects, i.e. via a unsupervised learning process. Due to the fact that the class labels are often not known before data analysis, for example, in case of being hard to assign class labels in large databases, clustering analysis is sometimes an efficient approach for analyzing such kind of data. To perform clustering analysis, similarity measures are often utilized to assess the distance between a pair of data objects based on the feature vectors describing the objects, in turn, to help assigning them into different object classes/clusters. There are a variety of distances functions used in different scenarios, which are really dependent on the application background. For example, cosine function and Euclidean distance function are two commonly used distance functions in information retrieval and pattern recognition[10]. On the other hand, assignment strategy is another important point involved in partitioning the data objects. Therefore, distance function and assignment algorithm are two core research focuses that attract a lot of efforts contributed by various research domain experts, such as from database, data mining, statistics, business intelligence and machine learning etc.

The main data type typically used in clustering analysis is the matrix expression of data. Suppose that a data object is represented by a sequence of attributes/features with corresponding weights, for example, in the context of Web usage mining, a usage data piece (i.e. user session) is modeled as a weighted page sequence. Like what we discussed above, this data structure is in the form of the object-by-attribute structure, or an n-by-m matrix where *n* denotes the number of data objects and *m* represents the number of attributes. In addition to data matrix, similarity matrix, where the element value reflects the similarity between two objects is also used for clustering analysis. In this case, the similarity matrix is expressed by an n-by-n table. For example, an adjacency matrix addressed in Web linkage analysis is actually a similarity/relevance matrix. In this work, we adopt the first data expression, i.e. data matrix to address Web usage mining and Web recommendation.

To date, there are a large number of approaches and algorithms developed for clustering analysis in the literature[11]. Based on the operation targets and procedures, the major clustering methods can be categorized as: Partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based methods, high-dimensional clustering and constraint-based clustering. Partitioning method is to assign *n* objects into *k* predefined groups, where each group represents a data segment sharing the highest average similarity in comparison to other groups. The well-known *k*-means is one of the most conventional partitioning clustering algorithms.

Hierarchical clustering provides an easily visualized way to modeling the underlying relationships among the data objects. Hierarchical clustering can be considered an agglomerative approach, which suffers from the problem of one-way construction, that is, it can not be undone during the hierarchical tree construction procedure. Model-based methods hypothesize that there exists a model for each of the clusters, into which one data object is best fitted by measuring a density function. The density function that associates with the special distribution of the data helps to determine the cluster number and to assign the data objects into various clusters. For example, *Self Organizing Map* (SOM) based clustering is one of the model-based methods, which is to map a data object in a high-dimensional space into a low-dimensional (e.g. 2-D or 3-D) grid map via a neural network based algorithm. The SOM-based clustering algorithm is eventually to map the original data objects onto different data blocks/segments of the SOM grid and the locality of the data points indicates the visualized clustering information.

User Profile Algorithms: Web clustering is one of the mostly used techniques in the context of Web mining, which is to aggregate similar Web objects, such as Web pages or users session, into a number of object groups via measuring their mutual vector distance. Basically, clustering can be performed upon these two types of Web objects, which results in clustering Web users or Web pages, respectively. The resulting Web user session groups are considered as representatives of user navigational behavior patterns, while Web page clusters are used for generating task-oriented functionality aggregations of Web organizations. Moreover, the mined usage knowledge in terms of Web usage patterns and page aggregates can be utilized to improve Web site structure designs.

There has been a considerable amount of work on the applications of Web usage mining and recommender systems. For example, Mobasher et al proposed an aggregate usage profile technique to cluster Web user transactions into various usage groups by using standard clustering algorithms, such as *k*-means clustering algorithm[6]. On the other hand, an algorithm called PageGather was proposed by Perkowith and Etzioni to discover significant page segments, which were used to help Web designers to add an additional index page that not existed before to facilitate Web users to locate their interested contents quickly, by using a Clique (complete link) clustering algorithm[5].

In the context of clustering, computational costs is a major concerned issue suffering researchers due to the particular characteristics of Web data, e.g. the problems of the high-dimension and the sparsity nature of Web data. For example, it is difficult, sometimes, to simply apply a standard clustering algorithm on the Web usage data with millions of user sessions to derive a collection of Web pages, which is resulting in a tough computational task. The reason is that instead of using pages as dimensions, the user sessions must be treated as dimensions and clustering is performed on this very high dimensional space. To address there issues, dimensionality reduction techniques and alternative clustering algorithms are explored. Amongst these, *Latent Semantic Analysis* (LSA) is considered as an efficient dimensionality reduction algorithm with the latent semantic analysis capability, that is, the capability of discovering the hidden knowledge from Web data by taking the semantic property of data into consideration.

Latent Semantic Indexing (LSI), one kind of traditional LSA algorithms, is a statistical method, which is to reconstruct a co-occurrence observation space into a dimension reduced latent space that keeps the maximum approximation of the original space by using mathematical transformation procedures such as *Singular Value Decomposition* (SVD). With the reduced dimensionality of the transformed data expression, the computational cost is significantly decreased accordingly, and the problem of sparsity of data is handled well as well. Besides, LSI based techniques are capable of capturing the semantic knowledge from the observation data, while the conventional statistical analysis approaches such as clustering or classification are in lack of finding underlying association among the observed co-occurrence. In last decades, LSI is extensively adopted in applications of information retrieval; image processing, Web research and data mining, and a large amount of successes have been achieved. In this chapter, we aim to integrate LSI analysis with Web clustering processes, to discover Web user session aggregates with better clustering quality, in other words, this techniques is on the basis of combination of latent semantic analysis and Web usage mining.

Latent Usage Information Algorithm: In this section, we present an algorithm called *latent Usage Information* (LUI) for clustering Web sessions and generating user profiles based on the discovered clusters. This algorithm consists of two steps, the first step is a clustering algorithm, which is to cluster the converted latent usage data into a number of session groups; and the next step is about generating a set of user profiles, which are derived from calculating the centroids of the discovered session clusters.

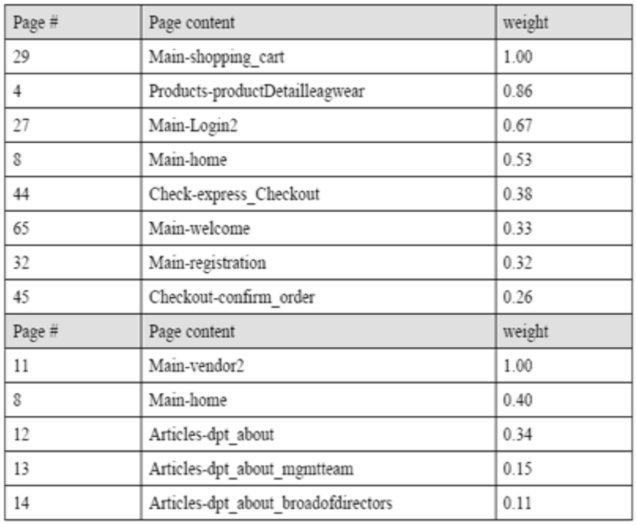
Building User Profile: As we mentioned above, each user session is represented as a weight-based page vector. In this way, it is reasonable to derive the centroid of the cluster obtained by the described clustering algorithm as a user profile. In this work, we compute the mean vector to represent the centroid.

Experimental Results: In order to evaluate the effectiveness of the proposed LUI algorithm, which consists of the Web clustering algorithm and the user profile generating algorithm, and evaluate the discovered user access patterns, we conduct experiments on two real world data sets and make comparisons with the previous work.

Results of User Profiles: We first utilize LUI algorithm to conduct Web usage mining on the selected two usage datasets respectively. We tabulate some results in below table-1 and table-2. In these tables, each user profile is represented by a sequence of significant pages together with corresponding weights. As we indicated before, the calculated weight is expressed in a normalized form, that is, the biggest value of them is set to be 1 while others are the relatively proportional values, which are always less than 1.

Table-1 depicts 2 user profiles generated from KDD dataset using LUI approach. Each user profile is listed in an ordered page sequence with corresponding weights, which means the greater weight a page contributes, the more likely it is to be visited. The first profile in table-1 represents the activities involved in online shopping behaviors, such as login, shopping cart, and checkout operation etc, especially occurred in purchasing leg-wear products, whereas the second user profile reflects the customers' concern with regard to the department store itself.

Table-1
Example of generated user profiles from KDD dataset



Page #	Page content	weight
29	Main-shopping_cart	1.00
4	Products-productDetailleagwear	0.86
27	Main-Login2	0.67
8	Main-home	0.53
44	Check-express_Checkout	0.38
65	Main-welcome	0.33
32	Main-registration	0.32
45	Checkout-confirm_order	0.26
Page #	Page content	weight
11	Main-vendor2	1.00
8	Main-home	0.40
12	Articles-dpt_about	0.34
13	Articles-dpt_about_mgmtteam	0.15
14	Articles-dpt_about_broadofdirectors	0.11

Table-2
Example of generated user profiles from CTI dataset

Page #	Page content	weight
29	Main-shopping_cart	1.00
4	Products-productDetailleagwear	0.86
27	Main-Login2	0.67
8	Main-home	0.53
44	Check-express_Checkout	0.38
65	Main-welcome	0.33
32	Main-registration	0.32
45	Checkout-confirm_order	0.26
Page #	Page content	weight
11	Main-vendor2	1.00
8	Main-home	0.40
12	Articles-dpt_about	0.34
13	Articles-dpt_about_mgmtteam	0.15
14	Articles-dpt_about_broadofdirectors	0.11

Analogously, some informative findings can be obtained in table-2, which is derived from CTI dataset. In this table, three profiles are generated: the first one reflects the main topic of international students concerning issues regarding applying for admission, and the second one involves in the online applying process for graduation, whereas the final one indicates the most common activities happened during students browsing the university website, especially while they are determining course selection, i.e. selecting course, searching syllabus list, and then going through specific syllabus. Looking at the generated user profile examples, it is shown that most of them do reflect one specific navigational intention, but some may represent more than one access themes.

Quality Evaluation of User Session Clusters: When the user session clustering is accomplished, we obtain a number of session clusters. However, how to assess the quality of the obtained clusters is another big concern for us in Web usage mining. A better clustering result should be that the sessions within the same cluster aggregate closely enough but keeping far from other clusters enough. After completing user session clustering, the next goal is to evaluate the quality of the generated clusters. In order to evaluate the quality of clusters derived by LUI approach, we adopt one specific metric, named *Weighted Average Visit Percentage* (WAVP)[6]. This evaluation method is based on assessing each user profile individually according to the likelihood that a user session which contains any pages in the session cluster will include the rest pages in the cluster during the same session.

From the definition of WAVP, it is known that the higher the WAVP value is, the better the quality of obtained session cluster possesses.

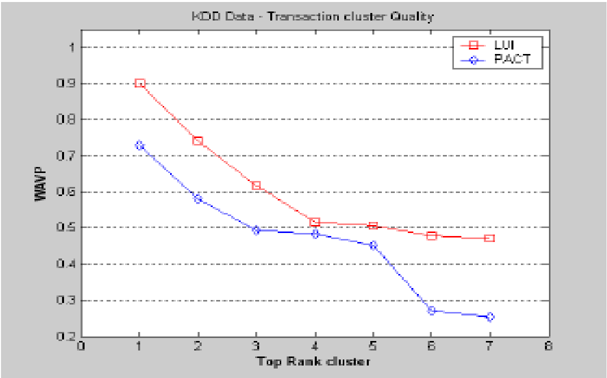


Figure-2
User cluster quality analysis results in terms of WAVP for KDD dataset

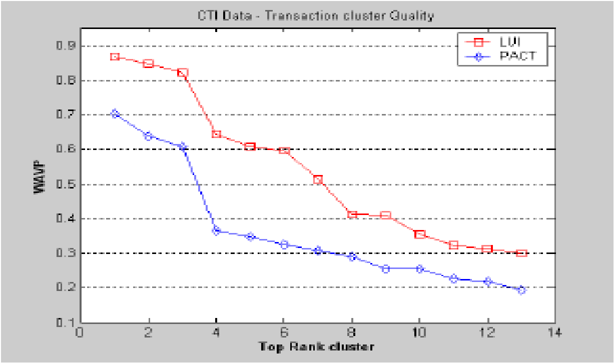


Figure-3
User cluster quality analysis results in terms of WAVP for CTI dataset

To compare the effectiveness and efficiency of the proposed algorithm with existing algorithms, here we use the PACT algorithm. We conduct data simulations upon two real world datasets by using these two approaches. Figure-2 and figure-3 depict the comparison results in terms of WAVP values for KDD and CTI datasets with PACT respectively. In each figure, the obtained user profiles are arrayed in a descending rank according to their WAVP values, which reflect the quality of various clustering algorithms. From these two curves, it is easily concluded that the proposed LUI-based technique overweighs the standard *k*-means based algorithm in term of WAVP parameter.

This is mainly due to the distinct latent analysis capability of LUI algorithm. In other words, LUI approach is capable of capturing the latent relationships among Web transactions and discovering user profiles representing the actual navigational patterns more effectively and accurately.

Related Work and Discussion: In the context of Web usage mining, there are two types of clustering methods performed on the usage data Web transaction clustering and Web page clustering[7]. One successful application of Web page clustering is the adaptive Web site. For example, the algorithm called PageGather[5] is proposed to synthesize index pages that do not exist initially, based on partitioning Web pages into various groups. The generated index pages are conceptually representing the various access interests of users according to their navigational histories. Another example is that clustering user rating results has been successfully adopted in collaborative filtering applications as a data preparing step to improve the scalability of recommendation using *k*-Nearest-Neighbour (kNN) algorithm[11]. Mobasher et al utilize Web transaction and page clustering techniques, which is employing the traditional *k*-means clustering algorithm to characterize user access patterns for Web personalization based on mining Web usage data[6]. These proposed clustering-based techniques have been proven to be efficient from their experimental results since they are really capable of identifying the intrinsic common attributes revealed from their historic clickstream data. Generally, these usage patterns are explicitly captured at the level of user session or page. They, however, do not reveal the underlying characteristics of user navigational activities as well as Web pages.

For example, such discovered usage patterns provide little information of why such Web transactions or Web pages are grouped together, and latent relationships among the co occurrence observation data have not been incorporated into the mining processes as well. Thus, it is necessary to develop LSA-based approaches that can reveal not only common trends explicitly, but also take the latent information into account implicitly during mining. An algorithm based on Principal Factor Analysis (PFA) model derived from statistical analysis, is proposed to generate user access patterns and uncover latent factors by clustering user transactions and analyzing principal factors involved in the Web usage mining[12]. Analogous, some studies are addressed to derive user access patterns and Web page segments from various types of Web data, by utilizing a so-called Probabilistic Semantic Latent Analysis (PLSA) model, which is based on a maximum likelihood principle from statistics.

Experimental Result: In this chapter, we have proposed a LSI-based approach, named LUI, for grouping Web transactions and generating user profiles. Firstly, we model the relationships among the co-occurrence observations (i.e. user sessions) into a usage data model in the form of a session-page matrix. Then, a dimensionality reduction algorithm based on the SVD algorithm has been employed on the usage matrix to capture the latent usage information for partitioning user sessions. Based on the decomposed latent usage information, we propose a *k*-means clustering algorithm to generate user session clusters. Moreover, the discovered user groups are utilized to construct user profiles expressed in the form of a weighted page collection, which represent the common usage pattern associated with one specific user access pattern. The constructed user profiles corresponding to various task oriented behaviors are represented as a set of page-weight pairs, in which each weight reflects the significance contributed by the page. Experiments have been conducted on two real world datasets to validate the effectiveness and efficiency of the proposed LUI algorithm. Meanwhile, an evaluation metric is adopted to assess the quality of the discovered clusters in comparison with existing clustering algorithms. The experimental results have shown that the proposed approach is capable of effectively discovering user access patterns and revealing the underlying relationships among user visiting records.

Clustering-Based User Profiling Techniques in Gait Pattern Mining: In the previous chapters, we intensively discuss Web usage mining for discovering user access patterns, and for predicting user navigational preferences and recommending the customized Web contents to Web users. During this procedure, clustering-based user profiling techniques (CBUP) plays an important role for usage knowledge discovery due to the capability of capturing the latent aggregate nature of co-occurrence observations.

In addition to its application in the area of Web information processing, CBUF can also be applied into a wide range of knowledge discovery and management domains, for example, in biomedical or health knowledge discovery and management fields, CBUF is usually used to create various typical characteristics to represent specific patient groups, which can be considered as pathological indicatives for various types of disorders or disease symptoms.

In this chapter, we aim to extend our developed methodologies and algorithms of CBUF from Web usage mining to a healthcare-based application, i.e. gait analysis, to investigate the hidden correlation among gait variables, discover normal and abnormal gait patterns in the form of gait variable vector and eventually explore the applicability of gait pattern mining in the diagnosis and analysis of human movement capability disorder. We carry out two case studies of gait pattern mining and give experimental results in this chapter. This chapter is organized as follows. In particular, we employ a SOM-based clustering algorithm to reveal gait patterns. Experimental analysis on two gait datasets is carried out to assess the proposed techniques.

Gait Data Model in Gait Pattern Mining: As for a biomechanical application of gait analysis, there is a variety of basic temporal distance parameters that are frequently used for modeling human walking, such as walking speed, stance/swing times. This may be due to the fact that the temporal-distance parameters are probably more fundamental for the purpose of gait analysis[13].

In this work, we simply exploit the specific two-dimensional temporal-distance parameters, i.e. stride length and step frequency/cadence to construct a gait data model. Both normal and pathological data relating to children's gait information for developing gait models were taken from[14]. In this model, the gait data is expressed as a two-dimensional feature vector matrix, in which each row represents a subject vector in terms of stride length and cadence parameters, whereas each column is corresponding to the selected gait variable.

Clustering-based User Profiling Algorithms for Gait Pattern Mining: Two types of clustering algorithms, i.e. *k*-means and hierarchical clustering are conducted to group gait data in terms of temporal-distance parameters. In *k-means* clustering analysis, we investigate the implementation of grouping the ambulation of neurologically intact individuals and those with CP into *k* subject categories, visualizing the separation layout of the grouped subject cluster and evaluating the clustering quality in terms of mean silhouette and mean square error.

In addition to *k*-means clustering, hierarchical clustering is also employed to reveal the possible grouping strategy for gait data from the viewpoint of hierarchy tree analysis. Meanwhile, construction of hierarchy tree and its corresponding visualization layout of clusters as well as centroids are plotted for comparing the clustering results derived by these two kinds of clustering algorithms[15].

User Profiling Algorithms for Web Recommendation: In previous section, we introduce a Web recommendation approach by identifying the user's task-oriented navigational distribution and incorporating it into the top-N weighted scoring scheme. Experiments conducted on the real world data sets have evaluated the proposed algorithm in terms of recommendation accuracy. The main idea of this approach is the use of the weights of pages within the dominant task space; however, it doesn't take the historical visits of other Web users into consideration. As a consequence, we aim to develop a Web recommendation algorithm via collaborative filtering techniques in this section. In particular, we propose two Web recommendation algorithms, which are called user profiling approaches based on two latent semantic analysis models.

The rest of this chapter is organized as follows: we first present two usage-based Web recommendation algorithms based on PLSA and LDA model respectively.

In order to evaluate the effectiveness of the proposed algorithms, comparison studies are carried out against existing Web recommendation algorithms.

User Profiling Algorithms for Web Recommendation: In this section, we present two usage-based user profiling algorithms for Web recommendation based on PLSA and LDA model respectively.

Recommendation Algorithm based on PLSA Model: Web usage mining will result in a set of user session clusters *SCL* = {SCL_1, SCL_2, $\_SCL_k$}, where each SCL_i is a collection of user sessions with similar access preferences. And from the discovered user session clusters, we can then generate their corresponding centroids of the user session clusters, which are considered as usage profiles, or user access patterns. The complete formulation of usage profiling algorithm is expressed as follows:

Given a user session cluster SCL_i, the corresponding usage profile of the cluster is represented as a sequence of page weights, which are dependent on the mean weights of all pages engaged in the cluster

$$up_i = \left(w_1^i, w_2^i, \cdots w_n^i\right)$$

where the contributed weight, w_j^i, of the page p_j within the user profile up_i is:

$$w_j^i = \frac{1}{\left|SCL_i\right|} \sum_{t \in SCL_i} a_{tj} \ ,$$

And a_{ij} is the element weight of the page *pj* in a user session s_t, s_t Є SCL_i. To further select the most significant pages for recommendation, we can use filtering method to choose a set of dominant pages with weights exceeding a certain value as an expression of user profile, that is, we preset a threshold µ and filter out those pages with weights greater than the threshold for constructing the user profile. Given w^i_j, then

$$w_j^i = \begin{cases} w_j^i, w_j^i > \mu \\ 0, otherwise \end{cases}$$

This process performs repeatedly on each user session cluster and finally generates a number of user profiles, which are expressed by the weighted sequences of pages. These usage patterns are then used into collaborative recommendation operations. Generally, a Web recommendation is to predict and customize Web presentations in a user preferable style according to the interests exhibited by individual or groups of users. This goal is usually carried out in two ways. On the one hand, we can take the current active user's historic behavior or pattern into consideration, and predict the preferable information to this specific user. On the other hand, by finding the most similar access pattern to the current active user from the learned usage models of other users, we can recommend the tailored Web content. The former one is sometimes called memory-based approaches, whereas the latter one is called model-based recommendations, respectively.

In this work, we adopt the model-based technique in our Web recommendation framework. We consider the usage-based user profiles generated in previous section as the aggregated representations of common navigational behaviors exhibited by all individuals in the same particular user category, and utilize them as a usage knowledge base for recommending potentially visited Web pages to the current user. Similar to the method proposed for representing user access interest in the form of an n-dimensional weighted page vector, we utilize the commonly used cosine function to measure the similarity between the current active user session and discovered usage patterns. We, then, choose the best suitable profile, which shares the highest similarity with the current session, as the matched pattern of current user[23]. Finally, we generate the top-N recommendation pages based on the historically visited probabilities of pages by other users in the selected profile.

Recommendation Algorithm Based on LDA Model: In this section, we present a user profiling algorithm for Web recommendation based on LDA generative model. LDA is one of the generative models, which is to reveal the latent semantic correlation among the co-occurent activities via a generative procedure. Similar to the Web recommendation algorithm proposed in the previous section, we, first, discover the usage pattern by examining the posterior probability estimates derived via LDA model, then, measure the similarities between the active user session and the usage patterns to select the most matched user profile, and eventually make the collaborative recommendation by incorporating the usage patterns with collaborative filtering, i.e.

Referring to other users' visiting preferences, who have similar navigational behaviors. Likewise, we employ the top-N weighted scoring scheme algorithm into the collaborative recommendation process, to predict the user's potentially interested pages via referring to the page weight distribution in the closest access pattern. In the following part, we explain the details of the algorithm.

Data, scope and limitation: Web data mining by means of web server access logs is becoming more powerful tool of collecting personal information from customers. Data gathered by this method can support implementation of complex relationship management strategies. Based on retrieved customer information consumers can be grouped in high lifetime valuable and low lifetime valuable, which would influence the services provided to them. Therefore the main benefits include: Increasing web information accessibility, understanding user's navigation behavior, user privacy, customer relationship, improving information retrieval, content delivery on web

Results and Discussion

The result and discussion of the paper will be clearly indicate that which techniques (Web Mining), approaches and architectures is the fastest of utilizing data mining methods to induce and extract useful information from Web information, services and goods online increases, data mining activities can expand rapidly allowing firms to retrieve highly personalized data about customers, which as well implies high privacy violations and concerns. Both marketers and users should follow privacy policy rules. Marketers should pay more attention to level of user trust and couple their data mining efficiency with respect to user privacy. In this dissertation, Kurt Thearling provides a business and technological overview of data mining and outlines how, along with sound business processes and complementary technologies, data mining can reinforce and redefine customer relationships.

Conclusion

The result of this paper will be clearly indicated that how web mining (in a broad sense, Data Mining applied to ecommerce) is applicable to improving the services provided by e-commerce based enterprises. Specifically, we first discussed some recent approaches and techniques used in data mining. We now present some ways in which web mining can be extended for further research. With the growing interest in the notion of semantic web, an increasing number of sites use structured semantics and domain ontologisms as part of the site design, creation, and content delivery.

Acknowledgement

First of all, we are sincerely pass our gratitude to our principal Dr. Vasudev Mishra and supervisor Dr. V. M. Thakare for their help, guidance and encouragement throughout the Research paper.

References

1. Shian-Hua Lin, Chi-Sheng Shih, Meng Chang Chen and et al., Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach. In Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia **(1998)**

2. Cooley R., Mobasher B. and Srivsatava J., Web Mining: Information and Pattern Discovery on the Word Wide Web, Technical Report TR 97-027, University of Minnesota, Dept. of Computer Science, Minneapolis, **(1997)**

3. Bray T., Measuring the Web, In Proceedings of the 5th Intl. WWW Conference, Paris, France **(1996)**

4. Chen M.S., Han J. and Yu P.S., Data Mining: An Overview from a Database Perspective, IEEE Transaction on Knowledge and Data Engineering, **8**, 866-833 **(1996)**

5. Perkowitz M. and Etzioni O., Adaptive Web Sites: Conceptual Cluster Mining. *In Proceeding of 16th International Joint Conference on Articial Intelligence*, 264-269, Stockholm, Sweden: Morgan Kaufmann, **(1999)**

6. Mobasher B. and et al., Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Data Mining and Knowledge Discovery **(2002)**

7. Mobasher B., Web Usage Mining and Personalization, in Practical Handbook of Internet Computing, *CRC Press*, **15,** 1-37 **(2004)**

8. Ellen, Spertus and ParaSite, Mining Structural Information on the Web, In proceedings of 6th International WWW Conference, April, **(1997)**

9. Wee-Keong Ng, Ee-Peng Lim, Chee-Thong Huang, Sourav Bhowmick, Fengqiong Qin. Web Warehousing: An Algebra for Web Information. In Proceedings of the IEEE Advances in Digital Libraries Conference, Santa Barbara, U.S.A., April, **(1998)**

10. Wang K. and Liu H., Schema Discovery for Semistructured Data. In Proceedings of International Conference on Knowledge Discovery and Data Mining, Newport Beach, AAAI, Aug. **(1997)**

11. Nestorov S., Abiteboul S. and Motwani R., Inferring Structure in Semi structured Data, In Proceedings of International Workshop on Management of Semistructured Data, **(1997)**

12. Zhou Y., Jin X. and Mobasher B., A Recommendation Model Based on Latent Principal Factors in Web Navigation Data, *In Proceedings of the 3rd International Workshop on Web Dynamics, ACM Press,* **(2004)**

13. Bhowmick Sourav S., Madria S.K., Ng W.K. and Lim E.P., Web Bags, Are They Useful in Web warehouse? In proceedings for 5th International Conference on Foundation of Data Organization, Japan, Nov. **(1998)**

14. O'Malley M.J. and et al., Fuzzy Clustering of Children with Cerebral Palsy Based on Temporal-Distance Gait Parameters, *IEEE Trans, ON Rehab. Eng*., **5(4),** 300-309 **(1997)**

15. Han J. and Kamber M., Data Mining: Concepts and Techniques **(2007)**

16. Bhowmick Sourav S., Ng W.K. and Lim E.P., Information Coupling in Web Databases, In Proceedings of the 17th International Conference on Conceptual Modelling (ER'98), Singapore, November 16-19, **(1998)**

17. World Wide Web Consortium. Document Object Model (DOM) Level 1 Specification. http://www.w3.org/TR/, **(1998)**

18. Wang K. and Liu H., Discovering Typical Structures of Documents: A Road Map Approach, ACM SIGR, August, **(1998)**

19. Florescu D., Levy A. and Mendelzon A., Database Techniques for the World Wide Web, A Survey, SIGMOD Record **(1998)**

20. Inman V.T., Ralston H.J. and Todd F., Human Walking, Baltimore **(1981)**

21. H. Vernon Leighton and Srivastava J., Precision among WWW Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos **(1997)**

22. Han J. and Fu Y., Discovery of Multi-level Association Rules. In Proceedings of International Conference on Very Large Databases, pages 420-431, Zurich, Switzerland, Sept, **(1995)**

23. Backman D. and Rubbin J., Web log analysis: Finding a Recipe for Success, **(1997)**

24. Pitkow J., In Search of Reliable Usage Data on the WWW, In Proceedings of the 6th International World Wide Web Conference, Santa Clara, California, April, **(1997)**

25. Baeza-Yates R. and Ribeiro-Neto B., Modern Information Retrieval, *Addison Wesley*, *ACM Press* **(1999)**

26. O'Conner M. and Herlocker J., Clustering Items for Collaborative Filtering. *In Proceedings of the ACM SIGIR Workshop on Recommender System*s, Berkeley, CA, USA: ACM Press **(1999)**

27. Madria Anjay, Bhowmick Sourav S., Ng W.K. and Lim E.P., Center for Advanced Information Systems, School of Applied Science, Nanyang Technological University, Singapore 639798 {askumar, p517026, awkng, aseplim}@ntu.edu.sg

28. Petrovskiy Ikhail, Faculty of Computational Mathematics and Cybernetics, Moscow State University Vorobjevy Gory, Moscow, Russia michael@cs.msu.su

29. By Juan D. Velásquez, PhD University of Tokyo, Assistant Professor, Department of Industrial Engineering University of Chile jvelasqu@dii.uchile.cl http://wi.dii.uchile.cl/

30. Heikki Mannila Nokia Research Center, P.O. Box 407 (Itamerenkatu 11) FIN-00045 Nokia Group, Finland Heikki.Mannila@nokia.com