Web Usage Mining

Chapter · July 2010

DOI: 10.1007/978-3-642-14461-5_6

3 authors:

Pablo Enrique Roman
University of Santiago, Chile
52 PUBLICATIONS **1,257** CITATIONS
SEE PROFILE

Gaston L'Huillier
Sudo Technologies, Inc.
35 PUBLICATIONS **462** CITATIONS
SEE PROFILE

Juan Domingo Velasquez
University of Chile
153 PUBLICATIONS **2,655** CITATIONS
SEE PROFILE

Chapter 6
Web Usage Mining

Pablo E. Román*, Gastón L'Huillier and Juan D. Velásquez

Abstract In recent years, e-businesses have been profiting from recent advances on the analysis of web customer behaviour. For decades experts have debated on ways of presenting the content or structure in a web site in order to captivate the attention of the web user in the web intelligence community. A solution to this could help boost sales in an e-commerce site. Web Usage Mining (WUM) is the extraction of the web user browsing behaviour using data mining techniques on web data. According to this, several models of data analysis have been used to characterize the Web User Browsing Behaviour. Nevertheless, outstanding techniques have recently developed in order to improve the conventional success rates for behavioural pattern extraction. In this chapter different approaches for WUM are presented, considering their main insights, results, and applications to web behaviour systems.

6.1 Introduction

The Internet has become a regular channel for communication, most of all for business transactions. Commerce over the Internet has grown to higher levels in recent years. For instance, e-shopping sales has been drastically increasing in the previous year, achieving a growth of 17% in 2007,generating revenue of $240 billion/Year

Pablo Román
Department of Industrial Engineering, University of Chile, República 701, Santiago, Chile, e-mail: proman@ing.uchile.cl

Gastón L'Huillier
Department of Industrial Engineering, University of Chile, República 701, Santiago, Chile, e-mail: glhuilli@dcc.uchile.cl

Juan D. Velásquez
Department of Industrial Engineering, University of Chile, República 701, Santiago, Chile, e-mail: jvelasqu@dii.uchile.cl

*

in the US alone [21]. This highlights the importance of acquiring the knowledge on how the Internet monitors customer's interaction within a particular web site.

One can compare this new technological environment using traditional marketing approaches, but the internet embraces new methods of determining consumers genuine needs and tastes. Traditional market surveys serve no purpose in reflecting the veracious requirements of customers who have not been precisely defined in the web context. It is well known that Web users are ubiquitous. In this sense, a marketing survey compiled on a specific location in the world does not carry clear statistical significance. However, online queries should improve this issue by requesting that each visitor answers as many focused question as they can [37]], but apart from predicting future customer preferences, online surveys can improve the effectiveness of the web site content strategy.

Web Usage Mining (WUM) can be defined as the application of machine learning techniques over web data for automatic extraction of behavioural patterns from web users . In this sense, the web usage patterns can be used for analyzing the web user preferences. Traditional data mining methods need to be pre-processed and adapted before employing over web data. Several efforts have been made to improve the quality of the resulting data, which are described in the data pre-processing chapter of this book. Once a repository of web user behaviour (Web Warehouse) is available [83], specific machine learning algorithms are applied in order to extract pattern regarding the usage of the web site. As a result of this process several applications can be implemented as adaptive web sites, such as recommender systems, and revenue management marketing amongst others.

For instance, the problems connected to the customization of web sites that are geared to improve sales are somewhat challenging. Recently, the one million dollar Netflix prize [46] has been contested after three years without a winner. NetFlix is an online DVD rental company, and its business is driven by online movie recommendations for customers based on its ratings of movie. The competition consists of improving the forecasting algorithm of the company for rating per users. The 2009 winning team used a modified linear stacked generalization algorithm for recommendations [89] only for improving the performance of the Netflix predictive algorithm by 10 percent. This is an example of how difficult it is to track or monitor the web user behaviour, but also it highlights the real life importance of such predictions. In this chapter the state of the art of WUM coupled with new trends that face the discipline is discussed.

6.2 Characterizing the Web User Browsing Behaviour

As described in [34, 49, 76, 78, 83], Web usage data can be extracted from different sources, from which web logs are considered as one of the main resources for web mining applications. The variety of different sources carries a number of complexities in terms of data pre-processing and furthermore these are associated with the incompleteness of each source. As a solution to this several pre-processing algo-

rithms have been developed [83]. Further sources like the hyperlink structure and the web content complement the extracted Log information, providing a semantic dimension of each user's action. Furthermore, in terms of the web log entries, several problems must be confronted. Overall, a web log in itself does not necessarily reflect a sequence of an individual user's documented access. Instead it registers every retrieval action but without a unique identification for each user.

6.2.1 Representative Variables

The web user browsing behaviour can be monitored by three kinds of data: the web structure, the web content and the web user session. The first is directly related to the environment. The third describes the click stream that each Web User performs during its visit to the web site.

- **The web structure:** A Web Site can be represented as a directed graph $G(N, V, T)$, consisting of a collection of n nodes $N = \{1, \ldots, n\}$ and vertices $V = \{(i, j)/$ *a web link point from i to j*$\}$ with text content $T = \{T_i\}$. A node i from G corresponds to a web page with text content T_i , the representation of the content will be described later. Two special nodes need to be individualized, as they do not correspond to any real page. This is because they represent the exit/entrance to the web site and each node consists of a link to the "exit or entrance" node. This representation has the advantage of explicitly including all transitions between nodes, which is useful for stochastic process descriptions.

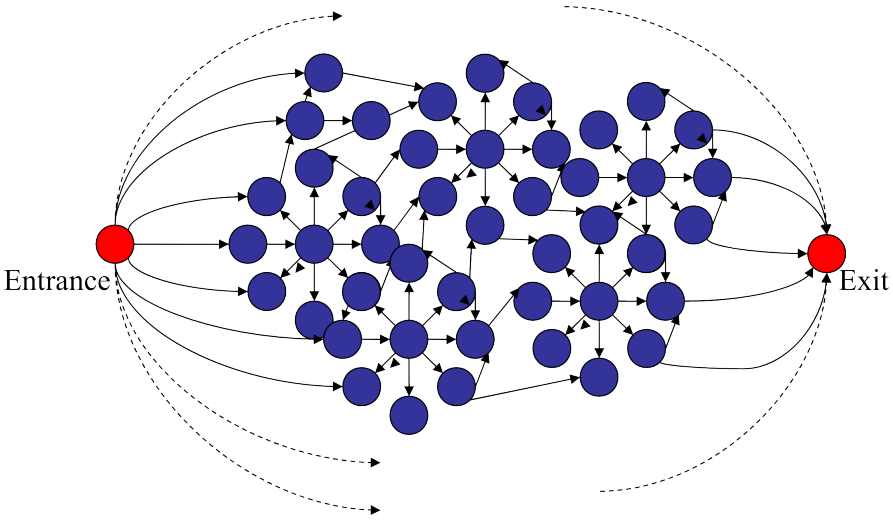

Fig. 6.1 Static graph web site structure representation.

Nevertheless, this description of a Web Site can be considered as a first approximation of the real hyperlink structure. The notion of a web page consisting of static content and unique URL can not fit the dynamical case. Web sites are being continuously updated, persistently modifying links and content, including those that depend on web user profile adaptive Web Site changes . On static pages with frames this concept is also challenged since the page is a composite. As stated in the Internet Report of China [22], the rate of web pages that moves from static to dynamic content is close to one.

Considering web 2.0 sites , the latter model seems obsolete. However, this approximation is valid considering some general circumstances. Informative sites that are updated on a regular periodic basis like those of newspapers can be represented under this graph representation within a definite period of time. Blogs can also be represented as such considering that the replies are increasingly added due to the fact that their nodes consist of post and reply.

More generally a time dependent graph structure of a variety of web objects and its associations can be defined for general purpose sites. For application, the analysis should be adjusted to the simpler model depending on the particular web site. Other representations of the structure of the web site are discussed in the data pre-processing Chapter.

- **The web content:** This corresponds to the web user perception semantic information of each visited page from a Web Site. On the earlier Internet this data corresponded mainly to the text content. Nowadays, Web 2.0 sites represent a much more complex picture. The web content is more dynamic and is constituted by a rich variety of media (text, images, video, embedded application, etc.). Web pages are composites of web objects that have semantic values. Several valuations of the semantic have been proposed and are revised in the chapter which describes semantic issues.

  Natural language processing for semantic extraction has been a large subject of study from when information retrieval systems began and it is still a large unsolved problem for reliable and automatic operational system [48]. Despite its limitation, some approximations to the problem have been proposed based on special representation of text and similarity measures at times have reasonable results. Instead of extracting the exact semantic, the notion of similar semantics between texts has demonstrated a more intelligent approach. The representations proposed are assigned to each term p in a page j, a weight ω_{pj} representing its semantic relevance. In this case, a column vector d_j of this matrix ω represents approximately the text semantic of the page j and row vector $t_p^\dagger$ represents the term p semanticly related to the set of documents. This approach is also termed "Bag of Word" because it does not take into account the semantic relations and syntax of phrases. In this way every equal word has the same importance independently of the context. Similar approaches can be extended to non-textual content like web object, where meta-data plays a fundamental role in representing the semantic.

  Furthermore, dynamic content implies time and user dependence of the semantic. As web applications become more complex than the standard representation

of the content the semantics become more inaccurate. Specific semantic context that group a variety of content must be tailored for each specific application.

- **The web user session:** Web User visits a Web Site represented by the browsing trajectorie that is categorised as a session [74]. A session s is a sequence $s = [(i_1, \tau_1), \ldots, (i_L, \tau_L)] \in S$ of pages $i_k \in N$ and time $\tau_k \in \mathbb{R}^+$ spend by a Web User. The size $L = \|s\|$ of a session corresponds to the number of nodes without considering the sink and source. In this representation the time associated with both the source and sink nodes and the duration of a session $\mathcal{T} = \sum_k \tau_k$ is the sum of all visitors times spent on the site.
  Nevertheless, if sessions are not explicitly given they must be reconstructed from other sources like web logs. When they are specially retrieved some privacy concerns [50, 36] arise that complicate its implementation. On the other hand, session retrieval from web logs have less relation to privacy issues since the data is stored anonymously. The process of extracting sessions has been reviewed in the chapter entitled data pre-processing.

Nevertheless, web data has some concerns. However, a further problem is associated with this data: the high dimensionality. Data mining algorithms suffer from the so called "curse of dimensionality" phenomenon. Over the years the processing of such data has been specialized as the Web Mining discipline, in particular when the purpose of such analysis is related to the behaviour of the Web User. In such cases it is called Web Usage Mining.

Also some evident problems exist that are connected with the web usage data. For example, the high diversity of some web pages; search engines that allow users to directly access some part of the web site; a single IP address with single server sessions; single IP address with multiple server sessions; multiple IP address with single server sessions; multiple IP address associated with a single visitor; and multiple agents associated with a single user session [83]. Additionally, a user's activation of the forward and reverse browser button is often not recorded in the web log because, in most cases, the browser retrieves the page from its own cache. A proxy server, acting as an internet web page cache serves to reduce network traffic, and can also capture web requests that are not recorded in a web log [20].

Browsing data has been recently considered in WUM, where the scroll-bar, select and save-as user interactions with the web site [78, 79]. Furthermore, semantic considerations have been proposed by different authors, where Plumbaum et al. in [61] uses the open standard of Microformats in order to add semantic information on the web page. This is a similar method as proposed in [70], for which JavaScript events (gathered with AJAX) are associated with key concepts in the portal, providing a context of such events and linking valuable usage information with the semantic of the web site.

As stated in [26], WUM presents different challenging problems in terms of the pre-processing of the usage data. Weblogs are larger in size, and both data volume and dimensionality, where sparse dataset representations of web users are needed to be transformed into a more accurate user behaviour representation. One of the problems associated to this lies in the high dimensionality, which results in computational complexity of the mining process. Secondly, the data scarcity results in the

mining algorithms having the ability to extract meaningful and interesting patterns in the user browsing behaviour.

Each WUM technique requires a model of user behaviour per web site in order to define a feature vector to extract behaviour patterns. Usually, the model contains the sequence of pages visited during the user session and some usage statistics, like the time spent per session, and pages viewed, amongst other information gathered. Here difficulties can be encountered when a page is loaded into a given web browser, the request for the web site objects can be logged separately, for which a series of page can be viewed associated with the same session.

All of the latter leads to the fact that web usage data requires different pre-processing techniques before analyzing the user behaviour [76]. However, one of the main tasks present in WUM is the determination of the web user sessions based on the web usage data collated from a given web site (sessionization). It is well known that strategies for sessionization can be classified as reactive and proactive [30].

Proactive sessionization strategies capture a rich collection of a web user's activity during the visit to a given web site. However, this practice is considered invasive, and even forbidden in some countries [74], or regulated by law to protect the user's privacy [43]. Examples of these methods include cookie oriented session retrieval [4], URL rewriting [18], and web tracking software, close to spyware, installed on the user's computer (or browser) to capture the entire session [50].

Reactive sessionization strategies have less privacy concerns because they are design to use only the web log entries' information, which excludes explicit user information [74]. However, a web log only provides an approximate way of retrieving a user's session for previously stated reasons. This reinforces the need to reconstruct a user's session from the information available (sessionization). Prior work on sessionization has relied on heuristics [3, 12, 74] which have been applied with a high degree of success on a variety of studies that include web user navigational behaviour, recommender systems, pattern extraction, and web site keyword analysis [83].

6.2.2 Empirical Statistics Studies about Web Usage

The human behaviour shows some predictive regularity on the averages, but to the contrary of the free will hypothesis. Some of those regularities are observed on the distributions of session in different kinds of web sites [29]. With the help of such regular statistics of the human behaviour Web Usage Mining can be tailored to fit those conditions. Several stochastic models have been theorized in order to mathematically explain this result [29, 80, 81], but nothing is related intrinsically to the physical phenomena. Some others models based on the neurophysiology of the decision making process have also been proposed [66].

Data mining processing on web data should show results that are in agreement with the observed universal probability distributions. As it was commented, web

user's actions on a web site follow regular patterns in probability distribution. This is additional information that can reduce the size of the feature space for machine learning algorithm. When such kind of reduction is available, algorithms have a narrow region for working resulting in better performance and accuracy. Nevertheless procedures must be adapted for fitting such statistical constrains [45, 57]. Understanding those statistics results in a better standard and quality for user [87].

Some important statistical empirical studies are summarized below.

- **Session Length Distribution**: Empirical studies over different web sites shows that the distribution of session size follows a common shaped function having an asymptotic heavy tail. Following [29] an Inverse Gaussian distribution ties in well with reality, and it was termed the universal law of surfing. In some work a Zipf distribution (power law) has been observed [44] reflecting a real session, but this distribution is used to approximate the Inverse Gaussian since its tails decay much slower than a Gaussian. Application of this kind of regularities enables the tuning up of systems like web crawler [2] and session retrieval from log file [13]. This web usage regularity has also been exploited for mining [45].
- **Information seeking behaviour**: While studies focus on algorithm for pattern extraction, few research relate to how web task users perform. Furthermore, the manner in which people seek information through the web can be classified through cognitive styles [88]. The studies differentiated two kind of web user: Navigators (17% of total of users) and Explorers (3%). The first, maintain consistency on the sequence of pages of visited pages. Navigators seek information sequentially and revisit the same sites frequently. The second have highly variable pattern of navigation. Explorers have a tendency to query web search pages frequently, revisit pages several times and browse a large variety of web sites. This kind of statistical study highlights the impact of classical navigational pattern extraction. Navigator will have the most influence on data regardless of consisting of 17 percent of total use. Nonetheless a whole skew distribution of cognitive styles have been reported [88], beginning with Navigators and ending with Explorers web users. The study of this distribution needs to be taken in account for further specialization of usage mining studies. A large study of cognitive information seeking has been investigated [32] showing that context can influence each particular web user behaviours. Others sources focus [38] on the statistics of the task that web users perform. Such taxonomy becomes associated with web user who perform transactions (e.g. reading emails 46.7%), Browsing (e.g. news reading 19.9%), Fact Finding (e.g. looking for whether 16.3%), Info Gathering (e.g. job hunting 13.5%) and a 1.7% non-classified. These four categories specify a simpler structure concerning the web usage.
- **Web User Habits**: Web user's habits have changed since the internet became more sophisticated. Nevertheless with current new web 2.0 applications web logs are becoming a much more intricate data source. Web users browsing behaviour are for ever changing, using fewer backtracking support tools such as the back button. However there is an increasing usage of parallel browsing with tabs and new windows.

- **The inter-event time for web site visit**: Similar heavy tailed distribution [56] has been measured on the time spent on pages throughout wide variety of the pages.

6.2.3 Amateur and Expert Users

Users can be grouped into two categories: experienced and inexperienced or "amateurs" [83]. The latter is unfamiliar with the process of accessing web sites and possibly dealing with web technology. Their behaviour is characterized by erratic browsing and sometimes they do not find what they are looking for. The former are users with web site experience and with some standard knowledge of web technology. Their behaviour is characterized by spending little time on pages with low interest and thus concentrating on the pages they are looking for, on which where they spend a significant amount of time. As amateurs gain experience, they slowly become experienced users who are aware of a particular web site's features. Therefore recommendations for change should be based on those users.

On the one hand, amateur users correspond to those unfamiliar with a particular web site and most probably with web technology skills [83]. Their browsing behaviour is erratic and often they do not find what they are looking for. On the other hand, experienced users are familiar with this or similar sites and have a certain degree of web technology skills. They tend to spend little time visiting low interest pages and concentrate on the pages they are looking for on which they spend a significant amount of time. As amateurs gain skills they slowly become experienced users, and spend more time on pages that interest them.

6.3 Representing the Web User Browsing Behaviour and Preferences

Regarding data mining purposes there are two kinds of structures that are used: features vectors that correspond to tuples of real number, and graph representation where numeric attributes are associated with nodes and relations. The most used representation corresponds to feature vectors with information about sessions, web page content and web site structure. Feature vector are employed for traditional web mining in an unsupervised a supervised fashion. Furthermore, Graph data structure is used for graph mining techniques or rule extraction.

6.3.1 Vector Representations

A session is a variable length of data structure that is not directly usable for most of the data mining algorithm. Web user activities can be extracted in several ways for the summarization of a session which is codified by mean of weighting ω_i the usage per page i. The vector $v = [\omega_i] \in V^n \subset \mathbb{R}^n$ has a dimension n corresponding to the number of different web pages. The cardinality of the set $|V| = m$ is equal to the number of collected session in the pre-processing phase. One of the most important information that this representation does not reflect, it is the sequence logic of each session. Despite this simplification this methodology has been used with success in several web mining applications.

Several methods exist to evaluate the weight ω_i. The simplest weighting schema corresponds to assigning a binary value $\omega_i \in \{0, 1\}$ represented if the page is used (1) or not (0) on this session [54]. More information can be incorporated extending the binary passage of the web page by the visit duration fraction. In this case the weight remains in the interval $\omega \in [0, 1]$ for normalization purposes. The fraction of time remaining in a web page is supposed to be an indicator of the quality and interest of the content [55]. Other weighting measures attempt at using other prior information about pages for measuring the significance of each page [55].

6.3.2 Incorporating Content Valuations

Hypermedia can be measured primarily by its text content. Natural Language Processing techniques consist of measuring text by means of different multidimensional representation. The most common valuation is the vector space model, where a document i is represented by a vector $m_i = [m_{ij}]$. Each component represents a weight for the importance of the word j in this document i. This model uses the "Bag of Word" abstraction, where any sentence structure is disregarded in favour of simple word frequencies. Furthermore, this semantic approximation has demostrated accurate results for data mining application. Several weighting schemas have been used with different results. The simplest is the binary weighting scheme where $m_{ij} = 1$ if the term j is present on the document i. The most used weighting scheme is the TF-IDF that combines the frequency of the term j in the document i and the frequency of document containing the term i. Recently the weight has been constructed with the help of a machine learning feature selection mechanism [42].

The text weighting scheme m_{ij} enriches the information provided by the feature vector of web usage. The visitor behaviour vector $v = [(m_i, \omega_i)]_{i=1}^n$, which each component represent the content and page importance.

6.3.3 Web Object Valuation

Nowadays web sites become highly dynamic applications. The visual presentation of a web page on a browser can not be identified correctly with a URL. A variable multiplicity of hypermedia could appear on browser presentation for the same URL. Nevertheless, embedded visual entities on web pages seem to be a more reliable concept. An object displayed within a webpage is termed a Web Object. Despite the complex semantic analysis of multimedia, metadata is used to define the Web Object within it (Figure 6.2).

Fig. 6.2 Some Web Object identified visually on browser and its XML representation.

Meta data that describes web objects constitute the information source for building the vector representation of content. The user's point of view is the principal research topic from which Web Object techniques has been developed. In this way the content and appearance of a web page is combined for processing. Different ways have been developed to describe web pages based on how the user perceives a particular page.

Web Object research has been carried in the following work: Web site Key objects identification [16], Web Page Element Classification [8], Named Objects [73] and Entity extraction from the Web.

6.3.4 Graph Representation

Graph mining uses graph theoretical construct and algorithm for discovering patterns in data [10]. On [23] web user trails are converted into a weighted graph using a similarity measure between session. Nodes correspond to sessions and arcs are labelled by the value of the similarity measure between both nodes. This similarity measure is used to overlap degrees between sessions, but any other measure can be adopted. The resulting structure is a direct representation of web user behaviour similarity.

6.3.5 The High Dimensionality of Representation

Usage data corresponds to a high dimensional representation. Considering that a medium sized web site contains thousand of pages and that around ten thousand terms exist, therefore the feature vector dimensionality correspond to at least 10^4 components. Automatic data mining methods based on similarity measure suffers from the "curse of dimensionality". This problem corresponds to the exponential grows of the size of the search space for data mining algorithm. Performance issues are one of the principal problems that could render the problem intractable. Recent studies reveal that distance based algorithm is affected since numeric difference between distance measure collapses in higher dimensional space. Similarity measure based algorithms (e.g. Clustering) are highly affected by this phenomenon.

Feature selection is a technique for data refinement that has been recently used for alleviating the higher dimensionality problem [31, 90]. A radical method of reducing the dimension of the feature vector is by way of a supervised approach to the text representation of web pages [63]. In this case an expert categorizes the semantic of the web pages which help to reduce the dimension of the feature vector.

6.4 Extracting Patterns from Web User Browsing Behaviour

As defined by Srivastava et al. in [76], "Web usage mining (WUM) is the application of data mining techniques to discover usage patterns from Web data, in order to understand the needs of Web-based application". Given this, one of the most challenging topics is the understanding and discovery of usage patterns from human users. Moreover, to analyze and predict human behaviour is its main characteristic, which is differentiated from Web Structure Mining and Web Content Mining, where techniques adopted are from a different nature and span between researchers and practitioners. It is relevant to point out that WUM is sometimes referred to as click stream analysis [76], considered as the analysis and extraction of underlying patterns from an aggregated sequence of page visits from a particular web site user navigation.

Interest in WUM is growing rapidly in the scientific and commercial communities, possibly due to both its direct application to web personalization and the increased complexity of web sites [83], and Web 2.0 applications [61]. In general, WUM uses traditional behavioural models, operations research and data mining methods (which will be intensively reviewed in this chapter) that deal with web usage data. However, some modifications are necessary according to their respective application domain, due to the different types of web usage data.

In general terms, two families of techniques have been used to analyze sequential patterns: deterministic and stochastic techniques. Each one of these techniques gathers different optimization methods, from data mining, operations research, and stochastic modelling, where different approaches have been adopted to compensate for the lack of analysis regarding some of these techniques present.

6.4.1 Clustering Analysis

Clustering user sessions can be used, essentially, for grouping users with common browsing behaviour and those interested in pages with similar content [76]. In both cases, the application of clustering techniques is straightforward when extracting patterns to improve the web site structure and content. Normally, these improvements are carried out at site shut-down.

However, there are systems that personalize user navigation with online recommendations about which page or content should be visited [60, 84]. In [35, 67], the *k*-means k-means clustering method is used to create navigation clusters. In [91], the usage vector is represented by the set of pages visited during the session, where the similarity measure considers the set of common pages visited during the user session.

Following previously stated reasoning, in [35], the pages visited are also considered for similarity with reference to the structure of the web site as well as the URLs involved. In [27] and [67], the sequence of visited pages is incorporated into the similarity measures additional to the page usage data. In [82], it is proposed that, despite the page sequence, the time spent per page during the session and the text content in each page is included in the similarity measure.

Also, clustering techniques such as Self-Organizing Features Maps (SOFM) have been used to group user sessions and information that could lead to the correct characterization of the web user behaviour [83]. In this context, Velásquez et al. in [85] proposed two feature vectors, composed by the information related to the text content from a web site, and a feature vector related to the usage, including information such as the time spend on the visited page by the web user. With this, the authors obtained the relevant words from the web site (or keywords), methodology which was extended by Dujovne et al. in [16] to determine the web site key objects.

Graph clustering techniques are applied to graph web data representation; through which [23] the time spent on each page was considered for similarity calculation and

graph construction. The representation of a session is created by using a similarity measure on the sequence of visited pages. A weighted graph is constructed using a similarity measure, and this in turn is achieved by employing an algorithm of sequence alignment [23]. This representation is further processed using a graph clustering algorithm for web user classification. Association rules can also be extracted [14].

Recently, Park et al. in [58], refers to the relevance of clustering in WUM that aims to find groups whose common interests and behaviour are shared. In this work, the question is whether sequence based clustering performed more effectively than frequency based, and thus acquired the best results a sequence based fuzzy clustering methodology proposed by the authors. Likewise, another recent clustering based methodology in WUM was proposed by Rios et al. in [63], where a semantic analysis was developed by considering a concept-based approach for off-line Web site enhancements. Here, the main development was the introduction of concepts into the mining process of WUM, carried out by a hybrid method based on the evaluation of web sites using Nakanishi's fuzzy reasoning model, and a similarity measures for semantic web usage mining.

6.4.2 Decision Rules

Decision rule induction is one of the classification approaches widely used in web usage mining [76]. Its practical results are sets of rules that represent the users' interests. In WUM, the association rules are focused mainly on the discovery of relations between the pages visited by the users [52]. For instance, an association rule for a Master and Business Administration (MBA) program is mba/seminar.html mba/speakers.html, showing that a user who is interested in a seminar tends to visit the speaker information page. Based on the extraction rules, it is possible to personalize web site information for a particular user [53].

6.4.3 Integer Programming

On terms of deterministic sessionization, Dell et al. in [13] presents a session reconstruction algorithm based on integer programming. This approach presents different advantages to address linking structure constraints presented by the time sequence of the web log entries, finding the best combinations of path that fulfill these constraints. Recent advances on integer programming and optimization [5], allows solving hard combinatorial problem with an acceptable timing. This method will be presented with further details in the data pre-processing chapter.

6.4.4 Markov Chain Models

Several statistical models for web surfing and web user behaviour have been developed in [6, 11, 17, 33, 75]. Here Markov models have been proposed for the modelling of the behaviour of the web user. In this cases, a web site is considered as an undirected graph of pages, in which the web user passes from one page to another [33, 64] within an estimated probability.

On the basis of a large number of users, and taking into consideration that the web site browsing is performed during a large period of time, it is possible to predict transition probabilities between pages. It can be considered that Web users have limited memory about visited pages, through which a transition probabilities can be considered independent of older transition probabilities. If transition probabilities are independent in more than k previous stages then the chain is known as a k-order Markov chain. Let X_o be the page visited at the step o and then a k-order Markov chain must have property presented in equation 6.1,

$$P(X_o|X_{o-1},\ldots,X_1) = P(X_o|X_{o-1},\ldots,X_{o-k}) \tag{6.1}$$

The latter expression represents a stochastic process with a k-step memory. A k-order Markov chain can be represented as a first-order Markov chain by re-labelling techniques [62].

Once the web user behaviour is modelled, and its flow probabilities determined, different levels of analysis and information can be extracted. Therefore, the prediction of the session size distribution, sequence of pages more likely to be visited, the mean time that a user can spend on the site or the ranking of the page that is most likely visited, are some of the possible outcomes that could be determined for the decision making process in a given web site.

Usually, a web site has a large amount of web pages and building transition matrices for the Markov model could be costly. For this reason, some authors [1] recommend the reduction of the dimensionality using clustering techniques over web pages. Then, the site transitions could be interpreted as web users changing from different clusters. The predictive power of the Markov chain has been studied by Sarukkai in [69], where the next browsing step of a web user is constructed taking the next link as the one with maximum probability. However, some authors proposed higher order Markov models to use for this task [15, 93], considering that lower order Markov models have been found with poor predictive power [9].

6.4.5 Mixture of Markov Models

Mixtures of Markov chains, a discrete set of Markov chains that represents different web user groups with different browsing behaviour [9, 71], has been proposed as an alternative for the web usage pattern extraction with emphasis on the different types of users a site might have. In this case, the independent relationship

with past behaviour is represented by an extended mixture coefficient $P(k) = \lambda_k$ into $P(X_o|X_{o-1}, \ldots, X_1) = \sum_{k=1}^{K} P(X_o|X_{o-1}, k)P(k)$, which represents K different browsing behaviour determined by further data mining techniques.

6.4.6 Hidden Markov Models

As an extension of the traditional Markov chain modelling, Hidden Markov Models (HMM) have been used to model the stochastic representation of the web site. However considering hidden states in the usage itself, web users have a underlying patterns in their behaviour [19, 72, 86]. Overall, this stochastic modelling tool has been used to determine frequent interest navigation patterns, combining web usage data and WCM techniques to build a predictive model.

6.4.7 Conditional Random Fields

Other approaches used to predict the web user behaviour are based on Conditional Random Fields (CRF) [41]. CRFs are a probabilistic framework generally used for the classification of sequential data. In the WUM context, Guo et al. in [25] aimed to predict all of the probable subsequent web pages for web users, comparing its results to other well known probabilistic frameworks, such as plain Markov chains and HMMs. Subsequently, in [24], an Error Correcting Output Coding (ECOC) of the CRF was proposed for the prediction of subsequent web pages on large-size web sites, extending previous development to a multi-class classification task for the web site prediction problem. It compared its results against single multi-label CRFs which outperformed the proposed method.

6.4.8 Variable Length Markov Chain (VLMC)

Another probabilistic framework developed for the web user behaviour modelling was proposed by Borges et al in [7], using as Variable Length Markov Chain (VLMC). There models are based on a non-fixed memory size, for which its usage in web browsing behaviour modelling is considerable. In [7], the main purpose for the researchers was to incorporate the dynamic extension of some web navigation sessions. They also aimed to extend the plain Markov chain method into a more general predictive tool. In this work, authors proved that the usage of such techniques increases the prediction accuracy, as well as the summarization ability of the Markov chain, an effective tool for the personalization of web sites, given that the web usage data is gathered by practitioners.

6.4.9 Biology Inspired Web User Model

A rather different approach from data mining consists of modelling the web user as a human operating a browser and performing decision about which hyperlink to click [66]. Neurophysiological stochastic model [65] has been studied from forty years ago establishing the stochastic evolution (Equation 6.2) of neuronal activity (Y_i) on the brain while the subject reaches a decision (Figure 6.3). Parameter of this model are $(\kappa, \lambda, \sigma)$ that are related to individual neuron configuration and the most important is I_i representing the likelihood of the choice i.

$$dY_i = (-\kappa Y_i - \lambda \sum_{j \neq i} f(Y_j) + I_i)dt + \sigma dW_i \tag{6.2}$$

Applying this theory to the browsing decision process leads to a complete and effective stochastic model of a web user. The model is based on the utility preferences per hyperlink. This utility function is modelled considering how each user is related to a text preference vector μ and a text similarity function between this vector and the hyperlink's text content representation. The user utility is higher if the hyperlink content vector is more similar to μ. A random utility model generates the choice a priori probability (I_i). The similarity measure used is the cosine between vector and the weighting scheme is the TF-IDF [66].

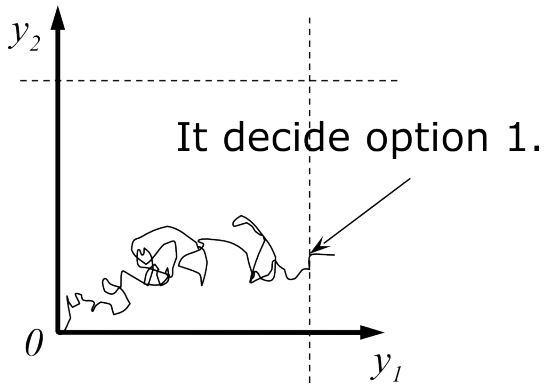

Fig. 6.3 Stochastic process for decision making. A decision is taken at the first hitting coordinate's time.

This new class of model has the capability to simulate the hyperlink aggregated demand and staying time on each page. This capability depends on the parameter adjustment using the available web usage data. A sub-product of the parameter fitting process is the probability distribution for text preferences $P(\mu)$. Furthermore predicting hyperlink is based on the content and structural modifications that can be obtained using stochastic simulation.

6.4.10 Ant Colony Models

Ant colonies have been used to learn the web usage [47]. The model is based on a simplification of a model based on the neurophysiology of the human decision making [66] and the random surfer [68]. The model called the "Ant Surfer" is where agents evolve like in the original random surfer model. The Ant Surfer start in a random page and continue browsing with probability p or return to the nest with probability $(1-p)$. The agent objective is foraging for information that is accumulated and its satiation is modelled with a threshold on the accumulated information utility. When the threshold is reach then it returns to the nest. The agent positioned on the page i select the hyperlink j to follows with probability P_{ij}, that correspond to the Logit model with utility given by the similarity measure between the ant text preference μ and the hyperlink text [66]. This model is applied to extract the web user preference and to predict web usage. Others models relate to other Markovian models for measuring the navigability of a web site [92] proposing similar web user models.

6.4.11 Matrix Factorization Methods

The NetFlix price [46] has been a corner stone evident to the magnitude of the Web Usage problem. A one million dollars price was announced for improving 10% the RMS error of the current NetFlix movie rental recommendation. The algorithm is based on matrix factorization which has been proven superior to similarity measure based techniques [40].The problem lies in mapping web usage vector and product vector (books, movies). In this model a join vector (u_l, p_k) of web user l behaviour vector u_l and the product k feature vector p_k and is linearly mapped to a factor space of dimensionality f.This problem stems from the family of singular value decomposition [59] where the linear mapping is partially known. The algorithm solves a minimum error square problem between known factor space rating values and user's and product data [77].

6.5 Application of Web Usage Mining

Web usage mining enables the analysis of the habits of a web user browsing in a web site. Furthermore knowing the user's interest and browsing behaviour can be used to improve a web site or build new web applications. The web usage mining could be used with automatic On-line algorithm, or in an Off-line fashion, supervised techniques. The general framework for the application of the web usage mining is the adaptive web sites [83]. The automatic personalization of the web site to the web users tastes, habits or marketing recommendation is only one of the techniques used on adaptive web sites for modifying or updating the content or structure.

6.5.1 Adaptive Web Sites

Adaptive Web Site are systems which adapt their content, structure and/or presentation of the Web Objects, to each individual user's characteristics, usage behaviour and/or usage environment [39].Adaptive sites provide users with both personalized and recommended services, and content according to the user's profile acquired by the system [28, 83]. The server load can be optimized since hyperlink demands can be forecasted and automatic balances could be performed. The topology of the web site can be modified to the web user interest. Different aspects of managing a web site are benefited from this technology. Marketing purposes have highly improved for adaptive sites. Usability trends are solved by means of specific user requirement.

6.5.2 Web Personalization

Web Personalization improves the web site structure based on the interaction with all visitors. Profiling is the principal processing that must be performed for those purposes. Using the profiling information, an automatic classification of web user should return the profile association with an objective (e.g. product). There are two kind of personalization [51, 83] based on the degree of conflict with the current semantic of the web site.

Tactical adaptation: It does not affect the overall structure of the web site as a semantic consistency is maintained and can be implemented by automatic systems. Such kinds of systems are autonomous and the whole design of the web site contemplates dynamic changes. Web sites like Amazon, NetFlix and others implement this kind of personalization.

Strategic adaptation: It must have the agreement of the owner of the web site, since the suggested changes are in conflict with the original orientation of the web site. Off-line recommendations are in general used for this kind of adaptation where owner's feedback is part of the component of the process [63].

6.5.3 Recommendation

Recommendation is also based on profiling and its objective is to retrieve a product that is most likely to be selected by the current web user. Furthermore, the profiling processing for web user's usage can be described from two different points of view depending on whether the profile is pre-established or is created on the run. The gen-

eral process is called “Filtering”, which is described depending on the orientation [40].

Content Filtering: Categories are created according to the nature of the Web User. The information about the user is retrieved from customer databases and associations with objectives like promotions or products are performed by a trained classifier.

Collaborative Filtering: The profile is created on the run based on the history of user’s interaction with the web site. The technique relates to the associations between products and relationship with users history. Some approaches relate to the product’s neighbourhood in which similar rating on other products is provided by other similar users. This technique is called Neighbourhood Method, where categories clusters are defined by the user past product rating. For instance a book entitled “Calculus” is visited in the neighbour of Authors like “Spivak” or “Apostol” as web users visit a web book selling site. Latent Factor Model is another approach based on recognizing factor variables that help to measure how useful a product is for a user. Once the factor is discovered, the importance of the produce is determined for a user. Matrix factorization methods have been used for implementing Latent Factor Models like in the 2009 winner NetFlix price [46].

6.6 Summary

Web Usage Mining has been studied for more than ten year with successful application. Nevertheless, vast quantity of new research in the area of applied behavioural sciences and other new trends such as matrix factorization methods have revolutionised the field of web mining. NetFlix price has demonstrated that traditional web usage mining techniques have little impact on real world issues. However, recent advances in this field have reaped promising results.

Acknowledgements This work was supported partially by the FONDEF project DO8I-1015, the *National Doctoral Grant* from *Conicyt Chile* and the *Web Intelligence Research Group* (`wi.dii.uchile.cl`) is greatly acknowledged.

References

1. Padmapriya Ayyagari and Yang Sun. Modeling the internet and the web: Probabilistic methods and algorithms. by pierre baldi, paolo frasconi, padhraic smith, john wiley and sons ltd., west sussex, england, 2003. 285 pp isbn 0 470 84906 1. *Inf. Process. Manage.*, 42(1):325–326, 2006.
2. R. Baeza-yates and C. Castillo. Crawling the infinite web: Five levels are enough. In *In Proceedings of the third Workshop on Web Graphs (WAW*, pages 156–167. Springer, 2004.
3. B. Berendt, A. Hotho, and G. Stumme. Data preparation for mining world wide web browsing patterns. *Journal of Knowlegde and Information Systems*, 1(1):5–32, 1999.

4. B. Berendt, B. Mobasher, M. Spiliopoulou, and J. Wiltshire. Measuring the accuracy of sessionizers for web usage analysis. In *Proc.of the Workshop on Web Mining, First SIAM Internat.Conf. on Data Mining*, pages 7–14, 2001.
5. Robert E. Bixby. Solving real-world linear programs: A decade and more of progress. *Operations Research*, 50(1):3–15, 2002.
6. José Borges and Mark Levene. Data mining of user navigation patterns. In *WEBKDD '99: Revised Papers from the International Workshop on Web Usage Analysis and User Profiling*, pages 92–111, London, UK, 2000. Springer-Verlag.
7. Jose Borges and Mark Levene. Evaluating variable-length markov chain models for analysis of user web navigation sessions. *IEEE Trans. on Knowl. and Data Eng.*, 19(4):441–452, 2007.
8. R. Burget and I. Rudolfova. Web page element classification based on visual features. *Intelligent Information and Database Systems, Asian Conference on*, 0:67–72, 2009.
9. Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, and Steven White. Visualization of navigation patterns on a web site using model-based clustering. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 280–284, New York, NY, USA, 2000. ACM.
10. D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.
11. Xin Chen and Xiaodong Zhang. A popularity-based prediction model for web prefetching. *Computer*, 36(3):63–70, 2003.
12. R. Cooley, B. Mobasher, and J. Srivastava. Towards semantic web mining. In *Proc. in First Int. Semantic Web Conference*, pages 264–278, 2002.
13. R.F. Dell, P.E. Román, and J.D. Velásquez. Web user session reconstruction using integer programming. In *Procs. of The 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 385–388, Sydney, Australia, December 2008.
14. Gül Nildem Demir, A. Sima Uyar, and Sule Gündüz Ögüdücü. Graph-based sequence clustering through multiobjective evolutionary algorithms for web recommender systems. In *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1943–1950, New York, NY, USA, 2007. ACM.
15. Xing Dongshan and Shen Junyi. A new markov model for web access prediction. *Computing in Science and Engg.*, 4(6):34–39, 2002.
16. L. E. Dujovne and J. D. Velásquez. Design and implementation of a methodology for identifying website keyobjects. In *Knowledge-Based and Intelligent Information and Engineering Systems, 13th International Conference, KES 2009, Santiago, Chile, September 28-30, 2009, Proceedings, Part I*, volume 5711 of *Lecture Notes in Computer Science*, pages 301–308, 2009.
17. Magdalini Eirinaki, Michalis Vazirgiannis, and Dimitris Kapogiannis. Web path recommendations based on page ranking and markov models. In *WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 2–9, New York, NY, USA, 2005. ACM.
18. F. Facca and P. Lanzi. Recent developments in web usage mining research. In *DaWaK*, pages 140–150, 2003.
19. Pedro F. Felzenszwalb, Daniel P. Huttenlocher, and Jon M. Kleinberg. Fast algorithms for large-state-space hmms with applications to web usage analysis. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *NIPS*. MIT Press, 2003.
20. Steven Glassman. A caching relay for the world wide web. In *Selected papers of the first conference on World-Wide Web*, pages 165–173, Amsterdam, The Netherlands, The Netherlands, 1994. Elsevier Science Publishers B. V.
21. K. Grannis and E. Davis. Online sales to climb despite struggling economy, 2008. According to Shop.org/Forrester Research Study.
22. K. Grannis and E. Davis. China internet network information center, 14th statistical survey report on the internet development of china 2009, 2009. According to http://www.cnnic.net.cn/uploadfiles/pdf/2009/10/13/94556.pdf.

23. Şule Gündüz and M. Tamer Özsu. A web page prediction model based on click-stream tree representation of user behavior. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–540, New York, NY, USA, 2003. ACM.
24. Yong Zhen Guo, Kotagiri Ramamohanarao, and Laurence A. Park. Grouped ecoc conditional random fields for prediction of web user behavior. In *PAKDD '09: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 757–763, Berlin, Heidelberg, 2009. Springer-Verlag.
25. Yong Zhen Guo, Kotagiri Ramamohanarao, and Laurence A. F. Park. Web page prediction based on conditional random fields. In *Proceeding of the 2008 conference on ECAI 2008*, pages 251–255, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press.
26. Tahira Hasan, Sudhir P. Mudur, and Nematollaah Shiri. A session generalization technique for improved web usage mining. In *WIDM '09: Proceeding of the eleventh international workshop on Web information and data management*, pages 23–30, New York, NY, USA, 2009. ACM.
27. Birgit Hay, Geert Wets, and Koen Vanhoof. Mining navigation patterns using a sequence alignment method. *Knowl. Inf. Syst.*, 6(2):150–163, 2004.
28. Wang Hongwei and Liu Xie. Adaptive site design based on web mining and topology. In *CSIE '09: Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering*, pages 184–189, Washington, DC, USA, 2009. IEEE Computer Society.
29. B. Huberman, P. Pirolli, J. Pitkow, and R. M Lukose. Strong regularities in world wide web surfing. *Science*, 280(5360):95–97, 1998.
30. Paul Huntington, David Nicholas, and Hamid R. Jamali. Website usage metrics: A re-assessment of session data. *Information Processing & Management*, 44(1):358–372, January 2008.
31. H. Hannah Inbarani, K. Thangavel, and A. Pethalakshmi. Rough set based feature selection for web usage mining. In *ICCIMA '07: Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, pages 33–38, Washington, DC, USA, 2007. IEEE Computer Society.
32. P. Ingwersen and K. Jirvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, first edition, 2005.
33. Søren Jespersen, Torben Bach Pedersen, and Jesper Thorhauge. Evaluating the markov assumption for web usage mining. In *WIDM '03: Proceedings of the 5th ACM international workshop on Web information and data management*, pages 82–89, New York, NY, USA, 2003. ACM.
34. Xin Jin, Yanzan Zhou, and Bamshad Mobasher. Web usage mining based on probabilistic latent semantic analysis. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 197–205, New York, NY, USA, 2004. ACM.
35. A. Joshi and R. Krishnapuram. On mining web access logs. In *Proc. of the 2000 ACM SIGMOD Workshop on Research Issue in Data Mining and Knowledge Discovery*, pages 63–69, 2000.
36. A. Juels, M. Jakobsson, and T.N. Jagatic. Cache cookies for browser authentication (extended abstract). In *SP '06: Proceedings of the 2006 IEEE Symposium on Security and Privacy*, pages 301–305, Washington, DC, USA, 2006. IEEE Computer Society.
37. A. Kausshik. *Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity*. Sybex, 2009.
38. M. Kellar. *An Examination of User Behaviour during Web Information Tasks*. PhD thesis, Dalhousie University, Halifax, Nova Scotia, Canada, 2007.
39. Constantinos Kolias, Vassilis Kolias, Ioannis Anagnostopoulos, Georgios Kambourakis, and Eleftherios Kayafas. Enhancing user privacy in adaptive web sites with client-side user profiles. In *SMAP '08: Proceedings of the 2008 Third International Workshop on Semantic Media Adaptation and Personalization*, pages 170–176, Washington, DC, USA, 2008. IEEE Computer Society.

40. Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
41. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
42. Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):721–735, 2009.
43. D. Langford. *Internet ethics*. MacMillan Press Ltd, 2000.
44. M. Levene, J. Borges, and G. Loizou. Zipf's law for web surfers. *Knowl. Inf. Syst.*, 3(1):120–129, 2001.
45. Jiming Liu, Shiwu Zhang, and Jie Yang. Characterizing web usage regularities with information foraging agents. *IEEE Trans. on Knowl. and Data Eng.*, 16(5):566–584, 2004.
46. Steve Lohr. A 1 million dollars research bargain for netflix, and maybe a model for others. *New York Times*, 2009.
47. P. Loyola, P. E. Román, and J. D. Velásquez. Colony surfer: Discovering the distribution of text preferences from web usage. In *Procs. Of the First Workshop in Business Analytics and Optimization (BAO)*, 2010.
48. C. D. Manning and H. Schutze. *Fundation of Statistical Natural Language Processing*. The MIT Press, 1999.
49. F. Masseglia, P. Poncelet, M. Teisseire, and A. Marascu. Web usage mining: extracting unexpected periods from web logs. *Data Min. Knowl. Discov.*, 16(1):39–65, 2008.
50. Viktor Mayer-Schonberger. Nutzliches vergessen. In *Goodbye privacy grundrechte in der digitalen welt (Ars Electronica)*, pages 253–265, 2008.
51. Alexander Mikroyannidis and Babis Theodoulidis. Heraclitus: A framework for semantic web adaptation. *IEEE Internet Computing*, 11(3):45–52, 2007.
52. Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Creating adaptive web sites through usage-based clustering of urls. In *KDEX '99: Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange*, page 19, Washington, DC, USA, 1999. IEEE Computer Society.
53. Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142–151, 2000.
54. Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Effective personalization based on association rule discovery from web usage data. In *WIDM '01: Proceedings of the 3rd international workshop on Web information and data management*, pages 9–15, New York, NY, USA, 2001. ACM.
55. Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Min. Knowl. Discov.*, 6(1):61–82, 2002.
56. H. Obendorf, H. Weinreich, E. Herder, and M. Mayer. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 597–606, 2007.
57. C. Olston and E.H. Chi. Scenttrails: Integrating browsing and searching on the web. *ACM Trans. Comput.-Hum. Interact.*, 10(3):177–197, 2003.
58. Sungjune Park, Nallan C. Suresh, and Bong-Keun Jeong. Sequence-based clustering for web usage mining: A new experimental framework and ann-enhanced k-means algorithm. *Data Knowl. Eng.*, 65(3):512–543, 2008.
59. A. Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD Cup and Workshop*, 2007.
60. Mike Perkowitz and Oren Etzioni. Towards adaptive web sites: conceptual framework and case study. *Artif. Intell.*, 118(1-2):245–275, 2000.
61. Till Plumbaum, Tino Stelter, and Alexander Korth. Semantic web usage mining: Using semantics to understand user intentions. In *UMAP '09: Proceedings of the 17th International*

Conference on User Modeling, Adaptation, and Personalization, pages 391–396, Berlin, Heidelberg, 2009. Springer-Verlag.
62. Sidney I. Resnick. *Adventures in stochastic processes*. Birkhauser Verlag, Basel, Switzerland, Switzerland, 1992.
63. Sebastián A. Ríos and Juan D. Velásquez. Semantic web usage mining by a concept-based approach for off-line web site enhancements. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 234–241, Washington, DC, USA, 2008. IEEE Computer Society.
64. Pablo E. Román and Juan D. Velásquez. Markov chain for modeling web user browsing behavior: statistical inference. In *XIV Latin Ibero-American Congress on Operations Research (CLAIO)*, 2008.
65. Pablo E. Román and Juan D. Velásquez. Analysis of the web user behavior with a psychologically-based diffusion model. In *Procs. Of the AAAI 2009 Fall Symposium on Biologically Inspired Cognitive Architectures*, page 72, Arlington VA, USA, 2009.
66. Pablo E. Román and Juan D. Velásquez. A dynamic stochastic model applied to the analysis of the web user behavior. In *The 2009 AWIC 6th Atlantic Web Intelligence Conference*, pages 31–40, Prague, Czech Republic, 2009. Invited Lecture.
67. T. A. Runkler and J. C. Bezdek. Web mining with relational clustering. *International Journal of Approximate Reasoning*, 32(2-3):217–236, February 2003.
68. Brin S. and Page L. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117, 1998.
69. Ramesh R. Sarukkai. Link prediction and path analysis using markov chains. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 377–386, Amsterdam, The Netherlands, The Netherlands, 2000. North-Holland Publishing Co.
70. Kay-Uwe Schmidt, Ljiljana Stojanovic, Nenad Stojanovic, and Susan Thomas. On enriching ajax with semantics: The web personalization use case. In Enrico Franconi, Michael Kifer, and Wolfgang May, editors, *ESWC*, volume 4519 of *Lecture Notes in Computer Science*, pages 686–700. Springer, 2007.
71. R. Sen and M. Hansen. Predicting web user's next access based on log data. *J. Comput. Graph. Stat.*, 12(1):143–155, 2003.
72. Jing Shi, Fang Shi, and HangPing Qiu. User's interests navigation model based on hidden markov model. In Guoyin Wang, Qing Liu, Yiyu Yao, and Andrzej Skowron, editors, *RSFDGrC*, volume 2639 of *Lecture Notes in Computer Science*, pages 644–647. Springer, 2003.
73. V. Snásel and M. Kudelka. Web content mining focused on named objects. In *(IHCI) First International Conference on Intelligent Human Computer Interaction*, pages 37–58. Springer India, 2009.
74. M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *Informs Journal on Computing*, 15(2):171–190, 2003.
75. Myra Spiliopoulou and Lukas Faulstich. Wum: A web utilization miner. In Paolo Atzeni, Alberto O. Mendelzon, and Giansalvatore Mecca, editors, *WebDB*, volume 1590 of *Lecture Notes in Computer Science*, pages 184–103. Springer, 1998.
76. J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 2(1):12–23, 2000.
77. Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. Major components of the gravity recommendation system. *SIGKDD Explor. Newsl.*, 9(2):80–83, 2007.
78. Yu-Hui Tao, Tzung-Pei Hong, Wen-Yang Lin, and Wen-Yuan Chiu. A practical extension of web usage mining with intentional browsing data toward usage. *Expert Syst. Appl.*, 36(2):3937–3945, 2009.
79. Yu-Hui Tao, Tzung-Pei Hong, and Yu-Ming Su. Web usage mining with intentional browsing data. *Expert Syst. Appl.*, 34(3):1893–1904, 2008.

80. A. Vazquez, J. Gama Oliveira, Z. Dezso, K.-I. Goh, I. Kondor, and Albert-Laszlo Barabasi. Modeling bursts and heavy tails in human dynamics. *PHYSICAL REVIEW E*, 73(3):036127, 2006.
81. Alexei Vazquez. Exact results for the barabasi model of human dynamics. *Physical Review Letters*, 95(24):248701, 2005.
82. J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. A new similarity measure to understand visitor behavior in a web site. *IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization*, E87-D(2):389–396, February 2004.
83. J.D. Velásquez and V. Palade. *Adaptive web sites: A knowledge extraction from web data approach*. IOS Press, Amsterdam, NL, 2008.
84. Juan D. Velásquez, Pablo A. Estévez, Hiroshi Yasuda, Terumasa Aoki, and Eduardo S. Vera. Intelligent web site: Understanding the visitor behavior. In Mircea Gh. Negoita, Robert J. Howlett, and Lakhmi C. Jain, editors, *KES*, volume 3213 of *Lecture Notes in Computer Science*, pages 140–147. Springer, 2004.
85. Juan D. Velásquez, Richard Weber, Hiroshi Yasuda, and Terumasa Aoki. A methodology to find web site keywords. In *EEE '04: Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04)*, pages 285–292, Washington, DC, USA, 2004. IEEE Computer Society.
86. Shi Wang, Wen Gao, Tiejun Huang, Jiyong Ma, Jintao Li, and Hui Xie. Adaptive online retail web site based on hidden markov model. In *WAIM '00: Proceedings of the First International Conference on Web-Age Information Management*, pages 177–188, London, UK, 2000. Springer-Verlag.
87. R. W. White. Investigating behavioral variability in web search. In *In Proc. WWW*, pages 21–30, 2007.
88. Ryen W. White and Steven M. Drucker. Investigating behavioral variability in web search. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 2007.
89. David H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.
90. Geraldo Xexeo, Jano de Souza, Patricia F. Castro, and Wallace A. Pinheiro. Using wavelets to classify documents. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:272–278, 2008.
91. Jitian Xiao, Yanchun Zhang, Xiaohua Jia, and Tianzhu Li. Measuring similarity of interests for clustering web-users. In *ADC '01: Proceedings of the 12th Australasian database conference*, pages 107–114, Washington, DC, USA, 2001. IEEE Computer Society.
92. Yuming Zhou, Hareton Leung, and Pinata Winoto. Mnav: A markov model-based web site navigability measure. *IEEE Trans. Softw. Eng.*, 33(12):869–890, 2007.
93. I. Zukerman, D. W. Albrecht, and A. E. Nicholson. Predicting users' requests on the www. In *UM '99: Proceedings of the seventh international conference on User modeling*, pages 275–284, Secaucus, NJ, USA, 1999. Springer-Verlag New York, Inc.