Chapter 11
Analysis of Usage Patterns in Large Multimedia Websites*

Rahul Singh and Bibek Bhattarai

Abstract User behavior in a website is a critical indicator of the web site's usability and success. Therefore an understanding of usage patterns is essential to website design optimization. In this context, large multimedia websites pose a significant challenge for comprehension of the complex and diverse user behaviors they sustain. This is due to the complexity of analyzing and understanding user-data interactions in media-rich contexts. In this chapter we present a novel multi-perspective approach for usability analysis of large media rich websites. Our research combines multimedia web content analysis with elements of web-log analysis and visualization/visual mining of web usage metadata. Multimedia content analysis allows direct estimation of the information-cues presented to a user by the web content. Analysis of web logs and usage-metadata, such as location, type, and frequency of interactions provides a complimentary perspective on the site's usage. The entire set of information is leveraged through powerful visualization and interactive querying techniques to provide analysis of usage patterns, measure of design quality, as well as the ability to rapidly identify problems in the web-site design. Experiments on media rich sites including the SkyServer – a large multimedia web-based astronomy information repository demonstrate the efficacy and promise of the proposed approach.

11.1 Introduction

The success of a website depends on users finding the information they seek. Understanding usage patterns is therefore a key step in optimizing web-site design and determining its usability. In general, user behavior for large websites is

*Jim Gray from Microsoft Research played a significant role in formulating some of the ideas involved in this chapter and in terms of overall advice and encouragement for this research. This work would not have happened without his participation.

R. Singh (✉) and B. Bhattarai
San Francisco State University, San Francisco, USA
e-mail: rsingh@cs.sfsu.edu

R. Chbeir et al., *Emergent Web Intelligence: Advanced Information Retrieval*, Advanced Information and Knowledge Processing, DOI 10.1007/978-1-84996-074-8_11,

complex and difficult to characterize. Advances in web-site design have further complicated this challenge due to two primary factors. *First* websites today are increasingly *media rich* in that, their content is expressed not just through text but through images, various forms of graphics, and even video and audio. Unlike text, the semantic content of media is much harder to discern algorithmically (the so-called signal-to-symbol-gap). This complicates analyzing the content of web-pages which in turn impacts reasoning about its usage. *Second*, many websites have started to support interaction modalities that extend beyond static browsing and link following. Examples of such user-data interaction modalities include Java-script enabled click-able images, click-able maps, and parametric and SQL-based database queries. For example, all of the aforementioned interaction modalities are supported in SkyServer [1], a media rich scientific (astronomy) website as well as more general websites such as CNN and Amazon. The presence of such diverse interaction modalities require development of new techniques to understand what information the user may have been looking for and determine how successfully/efficiently the information needs were satisfied.

The success of any website depends on users satisfying their information needs (finding the information they seek). Website usability can therefore be thought of as a measure of the ease with which users satisfy their information goals. Clearly, without the ability to survey the users, questions of usability and information goals can not be answered with certainty. Unfortunately, it is difficult to conduct such surveys once a site is live. Thus automated methods need to be developed that, given usage patterns, can estimate user information goals and measure how easily the structure of a site enables the fulfillment of these goals.

At the state-of-the art, techniques for usage analysis can be grouped into two broad categories; those that are based on the analysis of web-logs [7, 9, 11, 16, 19–21, 26, 28–32] and those that try to analyze the content of web-pages to model usage [2, 4–6, 12, 15]. Given a partially conducted transaction, web-log mining techniques seek to determine which page will be accessed next. The obvious way to approach this problem is to first extract patterns from the logs and then build a predictive model. Some of the strategies that have been proposed for model construction include the use of Markov models [9, 31], collaborative filtering [26], and various forms of association rule mining [20, 21, 30]. In contrast techniques that try to model usage through web-content analysis try to extract the information goal(s) for a browsing pattern based on web-content. Typically, such techniques do not use the information provided by the usage log. For example, they do not analyze the usage patterns by considering temporal information available in the logs.

Each of the aforementioned approaches has important limitations that impact their efficacy in real-world settings. For instance, usage-log mining provides information about how users are traversing the website. However, it cannot provide information either about the putative information goals underlying the user-behavior or about the extent to which the user information goals are satisfied. Consequently usage-log mining is, by itself, inadequate for assessing usability. A stark example of this can be obtained by considering the two actual users sessions from the SkyServer website [1] shown in Fig. 11.1. Log analysis shows that both user sessions followed

User Session 1:

P1: http://skyserver.sdss.org/dr2/en/

P2: http://skyserver.sdss.org/dr2/en/sdss/

P3: http://skyserver.sdss.org/dr2/en/sdss/data/data.asp

P4: http://skyserver.sdss.org/dr2/en/sdss/instruments/instruments.asp

User Session 2:

P1: http://skyserver.sdss.org/dr2/en/

P2: http://skyserver.sdss.org/dr2/en/sdss/

P3: http://skyserver.sdss.org/dr2/en/sdss/data/data.asp

Fig. 11.1 Two user sessions on the Skyserver. Web-log analysis can identify the similarity in the usage patterns while web-content analysis can provide cues to the underlying information goals that were being pursued

the similar browsing path, namely "P1, P2, P3, P4" for session-1 and "P1, P2, P3" for session-2. Log analysis also shows that user-1 left the site after visiting the page *instruments.asp*, while user-2 left from the *data.asp* page. But, log analysis can not address the question as to why both users followed similar paths but chose different pages to exit the site. Further it can not tell us what the possible information goals were which could have lead to the observed behavior or whether the user goals were satisfied at all.

In contrast to log analysis, content-based methods are based on the intuition that web content has a significant impact on the user navigation choices. These methods therefore seek to explain the user behavior based on the content of the pages visited. Content based techniques, such as [4, 5], require a fundamental model of user behavior. An important framework in this context is that of *information foraging theory* [23]. The basic idea of information foraging theory is that in a web site, the user makes traversal decisions looking for information that would satisfy his or her information goal. Thus, the traversal history and the content of pages visited can be representative of the user information need. However, by focusing solely on page content, methods based on this framework run the risk of missing important contextual information available through web-logs, such as load patterns, temporal sequencing of the usage patterns, source of requests etc.

In recent research, attempts have been made to analyze the web-content output by web servers with the goal of providing a summarized and high-level perspective of web usage [22]. While sharing many of our goals, especially in terms of information presentation, the research in [22] focuses on server analytics. It therefore discounts modeling the perceptual aspects of user-behavior and its potential impact on usage patterns. Research in the area of adaptive hypermedia [3] also shares many of our goals. Adaptive hypermedia systems try to adapt aspects of the system to user characteristics including user goals, knowledge, and background [18]. Such systems

typically include three important components: (1) a domain model, which specifies the conceptual design of the application, (2) a user model which contains information about the user, and (3) an adaptation model defining how the adaptation of the system is performed. Majority of adaptive hypermedia systems use an overlay model of user knowledge [18], where an estimation of user knowledge for each domain model is stored. An alternative is the historic model which utilizes user history of page visits. This assumes that past user behavior is a reliable indicator of future user actions. Another strategy is to build a model using data from a group of users and then use the model to make predictions about individual users. Different machine learning and statistical techniques have been utilized to build such models including [33]: linear models, TFIDF-models, Markov models, neural networks, clustering methods, rule induction, and Bayesian networks. One of the main problems in constructing such predictive models has been the collection of training data containing information about non-observable user characteristics such as user intention, user information needs, and interruptability. Experience sampling is one possible way to collect such information. In it, users are asked to reveal unobservable characteristics underlying their behavior during the course of activity. This information is subsequently used to build the predictive models. An overview of different experience sampling strategies and their comparison through user studies is provided in [17]. In yet other research, attempts have also been made to directly obtain information about information usage by user eye-tracking (in the specific context of perusing web-search results) [8]. Availability of additional information available through methods such as experience sampling or sensor-based monitoring of the physical user context can undoubtedly aid user-context modeling. However, deploying such solutions in real-world settings can be complicated either due to lack of access to users or due to privacy concerns.

The research presented in this chapter brings together and builds on many of the aforementioned ideas conceptually and technically. On the conceptual side, we combine web-page content with information from web-logs and contextual metadata about usage patterns. On the technical side, we consider both textual content as well as media-based content in web-pages during information goal determination as well as usage-flow modeling. Further, in our approach different forms of user-data interactions are accounted for (beyond users following static links). Finally, our research not only emphasizes the algorithmic aspects of usage analysis, but also demonstrates the role powerful visualization-query-exploration interfaces can play in utilizing human-machine synergy towards addressing this problem.

We begin this chapter by presenting an overview of the proposed approach in Section 11.2. This is followed in Section 11.3 by a description of the Skyserver, which is a large media-rich website and constitutes our primary experimental testbed. Section 11.4 outlines the proposed approach. Experiments and case studies are presented in Section 11.5. The chapter is concluded in Section 11.6 by reiterating the fundamental ideas behind our solution methodology and by outlining its broad applicability in designing and analyzing modern web-based information systems.

11.2 Overview of the Proposed Approach

Our approach involves two aspects. The first deals with the issue of user goal determination based on the observed usage patterns and multimedia page content and connectivity analysis. The second addresses the problem of presenting to web-designers and administrators, the considerable amount of contextual information related to usage patterns which is available from usage logs (such as access statistics, the distribution of geographical origin of activities, distribution of sessions durations, distribution of unique users over location or time, etc.) in a manner that is easy to interact with and assimilate. The integration of these perspectives has not found prominence in prior research efforts. However, this is crucial, since the identification of interesting/important usage patterns requires both modeling user behavior as well as the ability to interpret metadata related to usage patterns [5]. Thus the proposed approach brings together content and usage-pattern-based information goal determination with visualization and visual data mining of contextual metadata related to usage-patterns. By correlating these perspectives, user actions can be decomposed in terms of the following intuitive characteristics: *who, when, where*(pertaining to the spatial-temporal distribution of usage patterns and analyzed through visualization and visual mining of contextual metadata obtained from usage-logs), *what, how* (pertaining to user actions and obtainable through logs and content analysis), and *why* (related to the discerned user information goal – which provides a possible explanation of the user behavior). The interplay of these factors is graphically illustrated in Fig. 11.2.

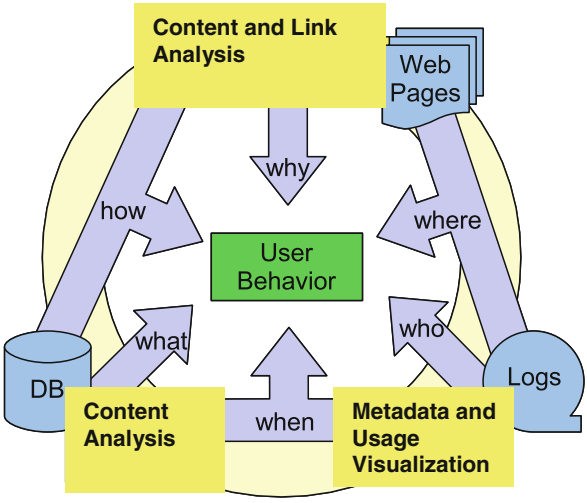

Fig. 11.2 Interplay of the three major concepts underlying the proposed method: content analysis (determining the core information content perused by the user), information goal determination (putatively explaining the user information need), and usage analysis and visualization (providing the analyst with analytical and contextual evidence on the usage patterns and the usability of the web-site)

We use ideas based on information foraging theory [23] to develop an explanatory model of user behavior. A key distinction of our work from prior research lies in the analysis of multimedia page content in estimating the information goals underlying the user behavior. Once the information goals are determined, an analysis of the linkage structure of the site provides the shortest path from the start page to the page(s) containing the information goal(s). If analysis of the usage logs show that multiple sessions diverge from this path, it may indicate a potential usability problem (such as critical links not clearly presented). The overall approach thus encompasses the following steps:

- Web content analysis to extract information goal related to the web session.
- Calculation of the overall user flow on the site for the extracted information goal. This provides a *simulated model* of how traffic on the site may be expected to behave for the specific information goals.
- Computation and comparison of the optimal (shortest) path with the path chosen by the user(s) to analyze usability and if/how the web-design may be improved to optimize access to information.
- Integrated visualization of the above information with contextual metadata extracted from the web logs.

11.3 Introduction to the Skyserver

The SkyServer website [1] constitutes the primary test-bed for development, testing, and validation of the proposed techniques. This website provides a large, media-rich real-world repository sustaining complex traffic and user-behavior patterns. In this section, we briefly introduce the reader to SkyServer and highlight its main characteristics.

SkyServer provides access to a large volume of astronomical data from the Sloan Digital Sky Survey (SDSS) [27]. The information is presented using text and a large number of images. Access to the information is provided through standard web browsers. The SkyServer website is designed to support a rich set of interactions between the user and the data [27], which include:

- Simple point-and-click interaction allows user to click on images of various different celestial object and retrieve data related to those objects.
- Text and GUI SQL web service interface where user can write their own query to access interact with SDSS database.
- Tools that let the user to enter astronomical information related to a particular object and retrieve its images and spectra.
- Skyserver is designed to support a diverse set of users starting from students learning astronomy at school level to scientists and professional astronomers. It should be noted that SkyServer is a *very* large website, offering views and data for over 80 million astronomical phenomena, totaling over one-and-a-half terabytes. The usage log data analyzed as part of this research is approximately 35 gigabytes and spans a timeline from May 2003 to October 2004.

11.4 Analysis of Usage Patterns

As the first step in the proposed approach, the content of the web-log is analyzed to derive contextual information that is important for understanding usage patterns. Dynamic content requests which culminate in an *Http GET* request or an *XMLHttp* request are recorded in the usage logs. Therefore these results can be reconstructed and analyzed. Web-log analysis begins with a data preprocessing step, where the data is scrubbed and validated. The preprocessing step is followed by *user delineation* and *user session definition*. These steps are elaborated below:

- *Definition of unique users:* Central to web usage analysis is the idea that users are discrete entities that exhibit (possibly multiple groups of) self-similar behavior in consuming web content. To categorize the behaviors, it is essential to identify each user. Prior work also explores the concerns regarding what constitutes a distinct user [25]. We define a unique user as having a distinct value for IP address and user-agent.
- *Definition of user sessions:* Studies have shown that user sessions are typically delimited by a timeout value of 25 min [24]. Based on these observations we use a timeout threshold of 30 min. For each discovered session, we cache the starting time of each session and the duration of the session as a whole.

These steps are followed by analysis of the web content. In this step we first extract the information stored in each page of the website. The process begins by constructing a sitemap for the website based on URL analysis. The site connectivity information is stored in an adjacency matrix. Subsequently, the text and media-based content in the site is analyzed. Using this information in the next stage, the putative information goals corresponding to the traversal pattern are identified. This is followed by the simulation of the user-flow. Finally this information is combined with metadata related to the usage patterns and presented using an interactive visualization interface. Using this interface the usability of the site can be analyzed.

11.4.1 Analysis of Text-Based Web-Content

The two most frequent media for most websites are text and images and therefore we focus on these in our analysis. Textual content is analyzed using a grammarless statistical method, which includes stemming and stop word filtering. First, the text-based content of each web page in the website is extracted and a vector of all unique terms present in the website constructed. Using the terms vector and the web pages connectivity information we construct the *term-page matrix*, TP_{TFIDF},

$$TP_{TFIDF}(i, j) = TFIDF(t_i, p_j) \tag{11.1}$$

In this matrix, the determination of the importance of a term t in a page p *is obtained* using its normalized TFIDF (Term-Frequency-Inverted-Document-Frequency)

value. This is a standard approach in text analysis, which essentially gives higher weight to more informative terms. The normalized TFIDF formulation used by us takes the length of the page into consideration for calculation of a term's importance and is defined as:

$$TFIDF = \left(\frac{tf}{N_{term}}\right) \times \log_2 \left(\frac{N_{page}}{df}\right) \tag{11.2}$$

In Eq. (11.2), *tf* is the frequency count of term in a given page, N_{term} is total number of terms in the page, N_{page} is total number of documents in the collection, and *df* is the frequency count of pages in which the term occurs and which *link to*, or *are linked from* the page of interest. This helps avoid the effect of unwanted terms by using a smaller, relevant document set as a background set.

11.4.2 Analysis of Image-Based Web-Content

After the text-based information content of a site has been captured, the next challenge is to represent its image-based content. In the case of web-pages the problem of determining the information corresponding to an image can be ameliorated by associating with the image, key snippets of proximal text. This brings up two sub-problems. First, the possible variability of semantics of an image (in terms of the text associated with the image) needs to be captured. This situation can arise when an image is used in multiple contexts. Second, the visual importance of an image has to be captured so that terms associated with highly prominent images receive greater weight as compared to terms associated with less prominent images. Solving these problems require the ability to describe and compare images.

In order to achieve these goals, we use color-texture analysis. Our approach uses the JSEG [10] color/texture analysis algorithm to identify textures within the image. Texture characterization is done with Grey-Level Co-occurrence Matrices (GLCM) [13]. We use eight vectors as the offset parameter for GLCM, and measure four statistical properties for each co-occurrence matrix: energy, entropy, contrast, and homogeneity. In addition, we generate a six-bit color histogram for each texture. Relative size, energy, entropy, contrast, homogeneity, and the color histogram are combined to create feature vector which is then used to describe each image in the web-site. Given two images represented by texture-color feature vector, their similarity is computed as the Pearson's distance between the vectors. A score of 1.0 indicates identical images and low scores indicate highly dissimilar images. If any specific image appears in two different pages, these pages have an image-semantic-based relationship. Consequently, the information from both pages contributes to the definition of the semantics associated with this image.

After image-based analysis is completed, key-terms from text proximal to images are used to describe the semantics associated with images while the term-frequency matrix captures the semantics of textual content of the website. Next, the information in the term-frequency matrix is combined with the image semantics to obtain a unified semantics representation of the entire information in the website. This is

done by re-weighting terms associated with images in the term-frequency matrix. The amount by which the weight values are adjusted is directly proportional to the size and complexity of the corresponding image. Such an adjustment is justified by the specificities of the human visual perception of images; an image with more texture (complexity) exhibits more information to the human eye than, for instance, an image of the same size but containing only single texture (such as an image with a uniform background of a single color).

11.4.3 Information Goal Extraction

For the extraction of the information goal, we first extract a list of terms from each page that is visited during the given session. The importance of each term in the list is calculated as summation of its TFIDF values across the pages visited in the session. Before summation, the TFIDF value of each term is multiplied by the importance value assigned to each page. For example, if we have a model where the final page in the session is accorded the greatest importance, then the terms appearing in the final page will be given greater weight as compared to appearing in other pages. Conversely, a model which weights all pages in a session equally can also be used. Finally, the term list is sorted and the 20 most important terms used as a summary of the user information goals. The specific number of terms used to summarize the user information goals is essentially a parameter which can be varied during analysis. Our choice of 20-terms was driven by the goals of obtaining a reasonable coverage of putative information goals without, at the same time, overwhelming the analysis with terms that may not be significant user goals.

To implement this idea, first a usage adjacency matrix U is constructed; if in a session a user visited the link from page i to page j then the matrix U is defined as shown in Eq. (11.3).

$$U(i,j) = \begin{cases} 1.0 : \textit{user visits page} \\ 0.0 : \textit{otherwise} \end{cases} \tag{11.3}$$

Next, the vector $I = \{I_p\}$ consisting of importance values corresponding to each page p in the site is constructed. In defining I, different weighting schemes may be used as described earlier. For instance, all pages can be weighted equally, or be weighted in incremental order (progressively increasing the importance value), or the final page can be weighted the highest (remaining pages weighted equally). Subsequently, the list of terms related to given session is obtained by multiplying TP_{TFIDF} with U and the vector I.

$$L = TP_{TFIDF} \times U \times I \tag{11.4}$$

Finally, the term weights are sorted to identify the top most informative terms.

11.4.4 Content-Based Usage Analysis

The first step in usage analysis involves computing the user flow through the website for a given information goal. Determining the user flow provides a probabilistic model of how other users with similar information goal behave, given the site structure and content. Our approach is based on the idea of information scent [23], which posits that users anticipate the information stored in distal page by looking at the text or graphical snippets (the *information scent*) present on the link pointing to the distal page. Consequently, given a specific information goal, links having information scent that strongly correlates with the information goal have a greater probability of being followed. The user flow is determined as follows:

- *Calculation of Information Correlation:* We calculate the correlation between users' information goal and information stored in a link by computing the normalized sum of the TFIDF value of all the terms that are present in *both* the URL and in the information goal as shown in Eq. (11.5). In cases where text is absent in the hyperlink, the title of the distal page is utilized in calculating the correlation. In Eq. (11.5), $C(l)$ is the correlation for link l, and t_{val} denotes the TFIDF value of term t in page P. Term t is also present in the information goal G.

$$C(l) = \frac{\sum_{tval} \forall t \in l, t \in G}{\sum_{i=1}^{n} t_i \forall t_i \in P} \tag{11.5}$$

- *User Flow Calculation:* The user flow is computed by simulating usage through an activation function A. The total percentage of users at a given time in a page depends on total information correlation value (IC) for all the links pointing to the page. The dampening factor α represents the fraction of users that can leave the website from any given page. The value of α can is determined based on site characteristics or by using the law of surfing [14].

$$A(t) = (\alpha \times IC \times A(t-1)) + E \tag{11.6}$$

In Eq. (11.6) E is the source activation vector and simulates users flowing through the links from the entry (or start) page of the usage pattern. The initial activation vector $A(1) = E$ and the final activation vector $A(n)$ gives the percentage of users in each node of the website after n iterations through the activation function.

- *Shortest Path Computation and Comparison:* Our underlying assumption is that the shortest path represents the most optimal (direct) path to the desired information goal. Thus, comparison of actual user paths with the shortest path provides cues to how well the links are organized in the website. For instance, repeated deviation of users from the shortest path may indicate usability issues such as an important link getting obfuscated due to design of the site. To compare a user-path with the optimal path we use a simple greedy strategy. We start the comparison from final page of the optimal path and seek to find a match starting

with the final page of the user-path and moving backwards to the start page. For every mismatch we assign a score of -1 and if a matching page is found we assign a score of $+1$ and mark the matching page. In the next iteration the page prior to the final page of the shortest path is considered and compared with the pages in the user path starting with the page prior to the matched page in the previous iteration and moving backwards. Again for each mismatched page, a score of -1 is assigned. The process is iterated until all pages in the shortest path have been sequentially compared with the pages in the user path. At the end the sum of scores over all the pages is calculated. The difference between this score and the shortest path length gives the measure of similarity between the user path and the optimal path. A score equal to length of shortest path means that the user path was identical to the shortest (most optimal path), while a score less than length of shortest path means user path differed from the shortest path.

As an example, consider the user-path and shortest path showed in Fig. 11.3, where the pages in the user-paths are labeled as *A, B, C, D* and *E*, and the pages in the shortest path labeled as *I, II* and *III* respectively. We weight the final page with the highest importance and therefore start from page *III* of shortest path and compare it with pages in users' path starting from page *E* to *A*. As page *III* matches with page *E*, we assign the score of $+1$ and mark page *E* as "matched". In next iteration we will take page *II* and start comparison from page *D* since page *E* is already matched. Page *II* matches with page *D*, we assign the score of $+1$ and mark page *D* as "matched". We now take page *I* and start comparing from page *C*. We find two mismatches at page *C* and page *B*, thus mismatch score of -1 will be assigned to each page. Finally page *I* matches with page *A* we assign $+1$ score to page *A*. The total score is then: $(+1) + (-1) + (-1) + (+1) + (+1) = 1$. This captures the fact that the user did not visit the direct link between page *A* and page *D*, but instead took the path page B– page C– page D.

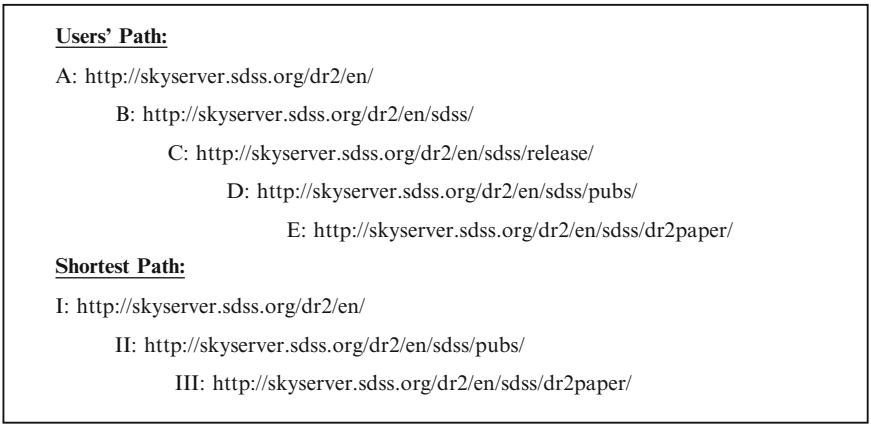

Fig. 11.3 Comparison between the path taken by a user and the optimal (shortest) path to the same information content

11.4.5 Contextual Metadata Extraction and Visualization

Obtaining a holistic understanding of usage patterns in a large website like the SkyServer requires combining the aforementioned analysis with contextual metadata associated with usage patterns. In our approach to this issue the basic principle lies in encouraging human-machine synergy by taking advantage of the human skills of domain expertise, contextual reasoning, pattern detection, hypotheses formulation, exploration, and sense-making. The process is facilitated through an integrated visualization-query-exploration interface (Fig. 11.4). The interface is *reflective*, which means that making a change in one of the components propagates the change to all other components. This interface also follows a direct manipulation paradigm. Perceiving a pattern of interest, a user can directly interact and explore the information. As a manipulation occurs in one of the information "views", its effects are appropriately reflected in the other views, thereby helping in understanding the relationships present in the data and its ultimate assimilation.

One of the features that the interface offers is the ability to directly correlate different aspects of information derived from the usage-logs. The correlation domains include: browsers (web agents), page, entry page, exit page, date, session duration, day of the week, month and hour. The correlation ranges include: *user sessions*, *user session duration*, *hits*, and *unique users*. The user can choose to chart any domain as the independent variable and any range as the dependent variable. For instance, a user can select an arbitrary slice of time (e.g. 2 weeks, 1 year, or 3 years) and

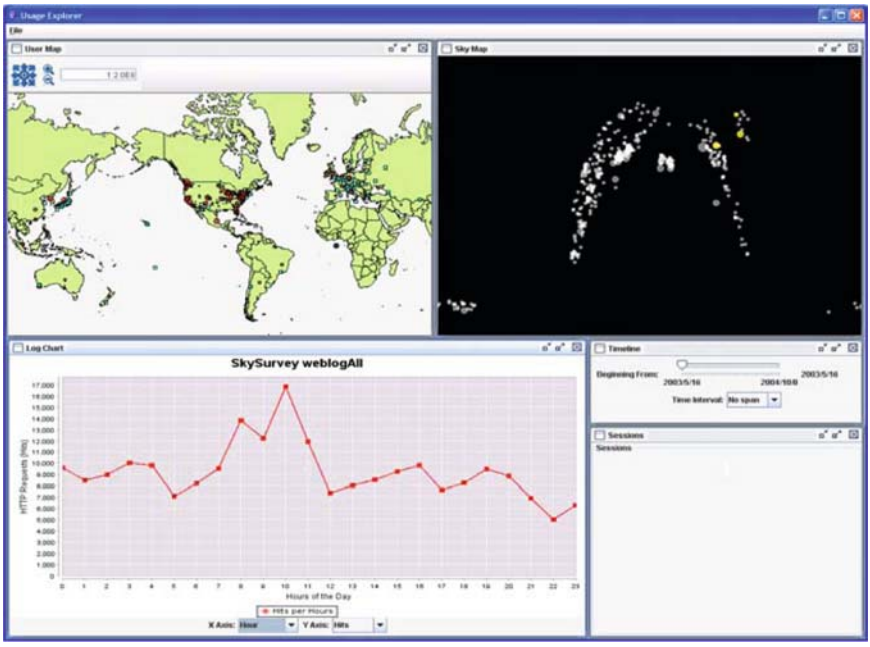

Fig. 11.4 Interface for visualization of contextual metadata

constrain the spatial information and log information accordingly. This mechanism can be used, for example to obtain a temporal distribution of the data as well as discern complex patterns such as recurring events or the influence of specific time periods on usage patterns. The visualization emphasizes the specific perspectives of *location* and *time* for analyzing the log data. Bearing in mind the cautions of [25] (with regards to user identity), we search for evidence of the location of the user by extracting information from various trustworthy web references, whenever possible. By visualizing this information on an interactive map, the user can explore the geographical distribution of the traffic (e.g. hits, sessions, visitors, etc.) is coming from. Further, by making the size of the dots proportional to the traffic volume, highly active sites/users can be easily identified.

A final component in the visualization shown in Fig. 11.4 (top right) is specific to astronomical data: in this part web pages about galaxies and stars are projected on a Cartesian map of the night sky. This provides a context-sensitive (astronomical) perspective of the data. This view is also reflected on the other views of the data.

In Figs. 11.5 and 11.6, we show further snapshots of the visualization interface Fig. 11.5 depicts the visualization of a user session. In this visualization, a directed graph is used to represent a subset of the web structure that is related to the user session being explored. For a given user session, the specific information being displayed include: (1) the path followed by user, (2) the shortest path as obtained after analysis with the proposed approach, and (3) the predicted user flow based on the specific information goal. Further, orange colored circles represent the web pages that have some relevance to the user goal and the red rectangular bars represent the total percentage of users that visit the page (based on the user flow). The orange directed lines (solid) linking the circles represents the direction of the user flow. Similarly, the green directed lines (dotted) and blue directed lines (dashed) represent the actual users' paths and system-computed shortest path, respectively. Pages that the user visits during the session, but are not present in the user flow graph, are represented by green squares. The same is true for blue triangles that represent the page that is present only in shortest computed path.

The visualization is interactive and dynamic, thus allowing researchers to explore the information and analysis. For instance, when a mouse pointer is moved over a square in the graph, information about the page (link information and thumbnail image of web-page) and percentage of user flow distribution in the page displayed as tool tip. The tool tip for the final page of a session also displays the match/mismatch score between user path and shortest path. This visualization provides a powerful tool that interfaces with the algorithmic aspects of our methodology and helps web designers rapidly analyze a user session to find, for example:

- The difference between the user path and the optimal path
- Page(s) where the user path diverges from the optimal path
- The predicted user flow for the information goal extracted

This allows web designers to rapidly identify the pages that have erroneous correlation value, driving users away from their goal. The system also provides the option to visualize users' paths, users flow graph, and the shortest path separately.

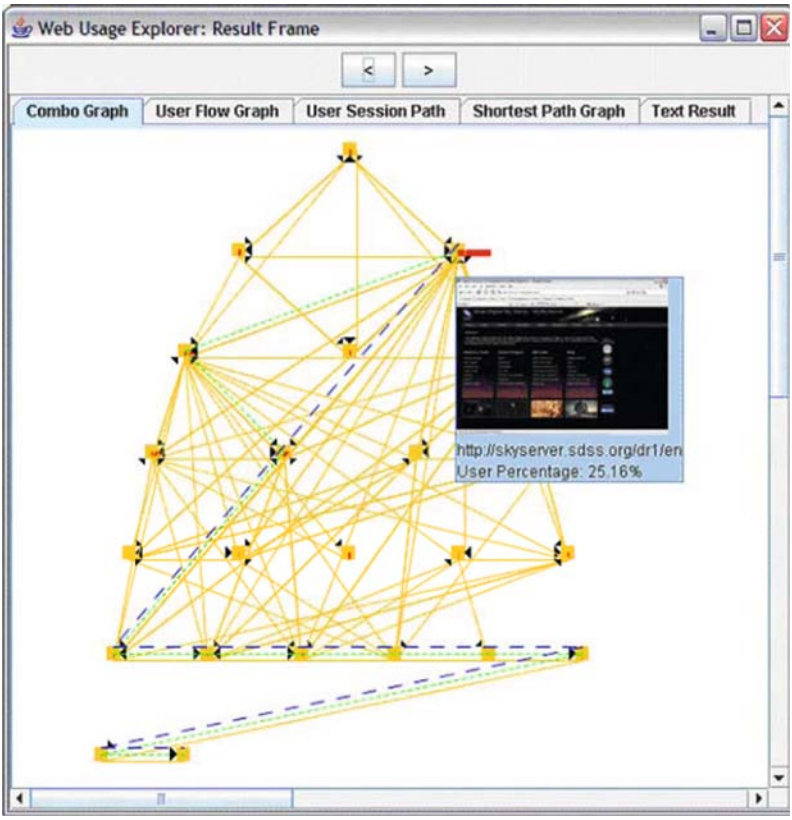

Fig. 11.5 Visualization support for usage analysis: The snapshot of the interface depicted in this figure shows the user session (Green dotted line), related shortest path (Blue dashed line) and user flow (Orange solid line). For a detailed discussion of the data depicted in this figure, the reader is referred to the case study in Section 11.5, where the user misses the optimal path to the information goal

The visualization framework also supports display and interaction with other types of contextual metadata relevant to the usage patterns. For instance, it is helpful to know where the majority of the website traffic comes from and which specific utilities/tools/functionality of the web-site are used the most. Such insights can be, for instance, help to address localization issues and system performance decisions such as scheduling system downtimes, or deciding where to place a redundant server for best workload distribution. Figure 11.6 presents a snapshot of one such visualization of the web traffic directed to a specific substructure of the Sky Server website (the search and query tools) with regards to the location of the internet service providers of the client. As this example illustrates, one of the search tools on Sky Server, *shownearest.asp*, is by far the most heavily used utility. The map also shows that the majority of the usage comes from the US costal regions (west coast not shown) and from around the Great Lakes.

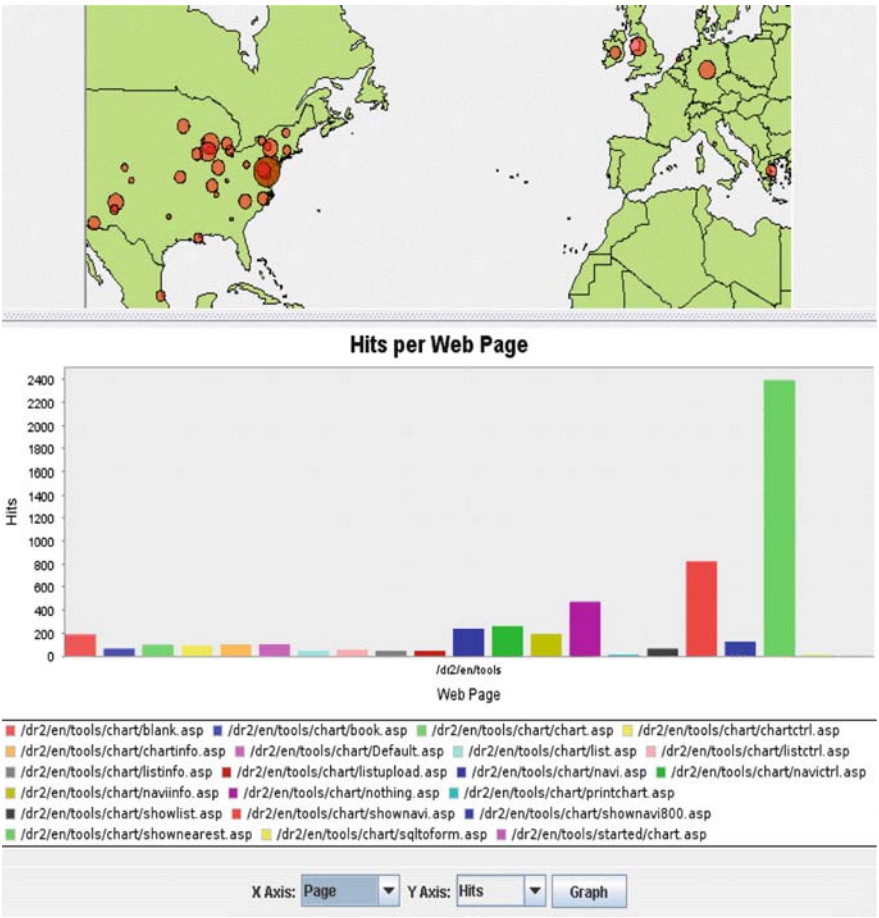

Fig. 11.6 Visualization of contextual metadata showing spatial distribution of the hits (top) and usage distribution for specific tools on the website (bottom)

11.5 Case Studies and Experiments

We begin with a case study involving a subset of the usage data from the SkyServer logs. The study starts with the data and visual interface shown in Fig. 11.4. Using the interactive interface, a visualization of the hits per month reveals that the month of April 2004 was very busy for the given usage-data sample; contributing 28.57% of the traffic between May 2003 and October 2004. We select, without loss of generality, traffic from Oxford, England which is shown on the map as a listed observatory. With these constraints, the system reveals four user sessions, one of which is long enough to have a meaningful information goal. This session starts from the index page, then browses to a spatial search page and executes three dynamic queries, each looking for a galaxy at specific points in the sky. The information goals predicted

Table 11.1 The comparative contribution of text-only content and text and image content to the user flow to select pages of the Skyserver website

Website subsections	Text only correlation (%)	Text/Image correlation (%)
Index page	25.16	25.29
Tools	33.06	28.81
Help	1.65	1.55
Traffic	0.00	4.10
Project	0.00	0.38

by our approach consist of the dynamic query results and static web page terms. Figure 11.5 shows the user flow (solid line), the user session path (dotted line), and the shortest path (dashed line). The visualization shows that the user missed the shortest path between the index and search pages of the site and that the search page was ultimately reached through the tool page. This is typical of a possible usability problem. Manually analyzing the corresponding pages, we find that the "tool-page" link visually dominates the "search-page" link on the index page. This causes users to follow the "tool-page" link and thus take a longer, indirect path. This case study illustrates how the proposed approach can be used by a designer to rapidly and interactively identify and correct problems in web-site design.

Using the parameters from this case study, we next present an experiment to evaluate the difference in the quality of results obtained by incorporating text-and-image information versus text-only information (see Table 11.1). When text-only information is used, no user flow is obtained for the (image rich) *traffic* and *project* subsections of the website, as they have no textual correlation with the information goal. In contrast, when *both* text and image-based information is considered, the activation function generates user flow to the traffic and project sections. This is due to the contribution of the image-based information to the information goal(s) and consequently the usage flow. It is interesting to note that the user flow simulation results obtained by incorporating both image and textual information, are supported by actual usage patterns from the logs and therefore constitute a more accurate model than what is possible with text dominated methods such as [4, 5].

Two related questions arise in the context of contrasting the multimedia content analysis strategy with a text-only approach. *First*, what is the impact of considering image content on information goals as compared with a text-only approach? *Second*, are there extreme cases where multimedia analysis can correctly identify information terms (goals) that could be totally missed in a text-only analysis?

We investigate these questions using data from the user sessions shown in Fig. 11.7. The results obtained from analyzing these user sessions are, based on our experience, representative for user sessions which involve media (image)-rich pages. The detailed results are presented in Table 11.2 and Fig. 11.8. For this specific user session, the terms *famous*, *place*, and *tool-title* are the top three goal terms in both text-only and text-and-image analysis. Table 11.2 shows that these terms have a reduced mean relevance using text-only analysis when compared with multimedia analysis. This is due to the fact that images contribute to increasing the relevance of these terms.

1: http://skyserver.sdss.org/dr1/en/tools
2: http://skyserver.sdss.org/dr1/en/tools/places
3: http://skyserver.sdss.org/dr1/en/tools/places/page1.asp

Fig. 11.7 A user session on the Skyserver used for analyzing the advantages of multimedia (text and image) analysis as opposed to text-only analysis

Table 11.2 Examples of changes in term relevance scores when multimedia (text and image) analysis is used instead of text-only analysis of the user session shown in Fig. 11.7

Term	Mean relevance for text-only analysis	Mean relevance for text-and-image analysis
Famous	0.0274	0.0369
Place	0.0272	0.0368
Tool-title	0.1000	0.1054

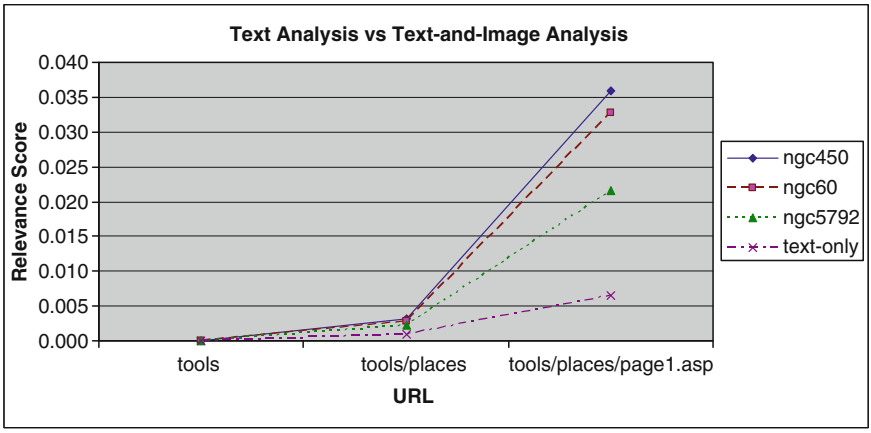

Fig. 11.8 Term relevance scores using the proposed multimedia content analysis approach versus text-only analysis of the pages

The extreme case is observed for the information goal terms *ngc450*, *ngc60* and *ngc5792*, which are names of galaxies prominently displayed on *page1.asp* (the third and final page visited in the session). The relevance scores of these terms are plotted across the session in Fig. 11.8 using both text-only and multimedia (text and image) analysis (proposed method). The reader may note that in the case of text-only analysis these terms have nearly negligible relevance values, since they only occur as image captions. In contrast, with the proposed method, the relevance values of these terms are significantly higher.

In the final experiment we perform a side-by-side comparison of information goal prediction for the user session shown in Fig. 11.9 using the proposed approach

1: http://skyserver.sdss.org/dr1/en/proj/challenges
2: http://skyserver.sdss.org/dr1/en/proj/challenges/hii
3: http://skyserver.sdss.org/dr1/en/proj/challenges/hii/characteristics.asp
4: http://skyserver.sdss.org/dr1/en/proj/challenges/hii/query.asp
5: http://skyserver.sdss.org/dr1/en/proj/challenges/hii/identifying.asp
6: http://skyserver.sdss.org/dr1/en/proj/challenges/hii/catalogs.asp

Fig. 11.9 A user session on the Skyserver used for comparing the proposed method with the IUNIS algorithm

Table 11.3 Comparison with the IUNIS algorithm

Proposed approach		IUNIS	
Terms	Mean relevancy (%)	Terms	Mean relevancy (%)
Region	14.44	Schema	0.19
Hii	14.26	Browser	0.16
Challenge	10.32	Query	1.04
Catalog	8.53	Dr1	0.00
Write	8.14	sdss	1.09

and the IUNIS algorithm [4]. It may be noted that both methods use a TFIDF-based measure to calculate term relevance and the term relevance scores are averaged over all pages in the session. Therefore the fundamental distinction is in how these methods account for non-textual information and how actual usage patterns are accounted in the analysis. Table 11.3 shows the top five putative information goals as determined by the proposed method and by IUNIS. The difference in the relevance of the terms obtained using the above methods is stark; the terms ranked as the top five terms by our approach have a significantly higher mean relevance score as compared with terms determined using IUNIS. Since the relevance of a term in a page is the TFIDF-based importance of the term, this indicates a weakness of IUNIS in that information goals identified by it can have low relevance (TFIDF scores). To understand why the IUNIS algorithm picks such terms we note that in highly interlinked websites like the Skysever, contents of nodes with large fan-in receive higher activation weights in INUIS even if they are not important in terms of the actual user goals. Thus terms with low relevancy-scores can get identified, incorrectly, to be important.

11.6 Conclusions

This chapter presents novel approach for usability analysis of large websites. We propose three fundamental extensions to the state-of-the art. First, we emphasize an integrative solution to this challenge that leverages and correlates information in

web-logs, the content of web pages, and contextual metadata to understand usage patterns. Second, we develop techniques that are capable of discerning information goals by taking into account information in the web-pages that may be expressed textually or through media such as images. Finally, we emphasize the role powerful visualization techniques can play, not only by enabling human machine synergy in analyzing complex patterns, but also by acting as the unification point around which the various analysis strategies can be brought together. Case studies and experiments conducted on real-world data from the SkyServer illustrate the efficacy of these ideas and their promise in developing a new generation of usage analysis strategies.

Acknowledgements The authors thank Mike Wong for his participation in parts of this project and to Jay Kim for help in formatting. This work was funded in part by a Microsoft unrestricted research grant to RS.

References

1. Sloan Digital Sky Survey project's website SkyServer: "http://skyserver.sdss.org/"
2. Blackmon M. H., Polson P. G., Kitajima M. Repairing Usability Problems Identified by the Cognitive Walkthrough for the Web. ACM CHI, pp. 497–504, 2003
3. Brusilovsky P., Adaptive Hypermedia, User Modeling and User Adapted Interactions, 11: pp. 87–110, 2001
4. Chi E. H., P. L. Pirolli, Chen K., Pitkow J. Using Information Scent to Model User Information Needs and Actions on the Web. ACM CHI, pp. 490–497, 2001
5. Chi E. H., Rosien A., Supattanasiri G., Williams A., Royer C., Chow C., Robles E., Dalal B., Chen J., Cousins S. The Bloodhound Project: Automating Discovery of Web Usability Issues using the InfoScent Simulator. ACM CHI, pp. 1323–1332, 2003
6. Cooley, R. The Use of Web Structure and Content to Identify Subjectively Interesting Web Usage Patterns. 2003 ACM Transactions on Internet Technology 3(2), pp. 93–116, 2003
7. Cooley, R., Mobasher B., Srivastava, J. Web Mining: Information and Pattern Discovery on the World Wide Web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), pp. 558–567, 1997
8. Cutrell, E. and Guan, Z., "What are you looking for?: an eye-tracking study of information usage in web search". Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 407–416, 2007
9. Deshpande M., Karypis G., "Selective Markov Models for Predicting Web Page Access", ACM Transactions on Internet Technology, 4(2), 163–184, 2004
10. Deng Y., and Manjunath B. S., Unsupervised segmentation of color-texture regions in images and video, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 8, pp. 800–810, 2001
11. Ding, C. and Zhou, J., "Improving website search with server log analysis and multiple evidence combination", International Journal of Web and Grid Services 3(2), pp. 103–127, 2007
12. Heer, J. and Chi, E. H. Identification of Web User Traffic Composition and Multimodal Clustering and Information Scent, Proc. Of the Workshop on Web Mining, SIAM Conference on Data Mining, pp. 51–58, 2001
13. Howarth P., Rüger S., Evaluation of Texture Features for Content-based Image Retrieval, Lecture Notes in Computer Science, Volume 3115, pp. 326–334, 2004
14. Huberman B., Pirolli P., Pitkow G., Lukose R., Strong Regularities in World Wide Web Surfing, Science, 280, pp. 95–97, 1998

15. Jin, X., Zhou, Y., Mobasher, B. Web Usage Mining Based on Probabilistic Latent Semantic Analysis. Proceedings ACM Special Interest Group on Knowledge Discovery and Data Mining, pp. 197–205, 2004
16. Joshi K., Joshi A., Yesha Y., Krishnapuram R., "Warehousing and mining Web logs", 2nd international workshop on Web information and data management, pp. 63–68, 1999
17. Kapoor, A. and Horvitz, E. "Experience sampling for building predictive user models: a comparative study" Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 657–666, 2008
18. Kravčík, M. and Gašević, D. "Adaptive hypermedia for the semantic web", Proceedings of the Joint international Workshop on Adaptivity, Personalization &Amp; the Semantic Web, pp. 3–10, 2006
19. Masseglia F., Poncelet P., Teisseire M., Using data mining techniques on Web access logs to dynamically improve hypertext structure, ACM SIGWEB Newsletter, vol. 8 no. 3, pp. 13–19, October 1999
20. Mobasher B., Dai H., Luo T., Nakagawa M., "Effective Personalization Based on Association Rule Discovery from Web-Usage Data", ACM Workshop on Web Information and Data Management, pp. 9–15, 2001
21. Nanopoulos A., Katsaros D., and Manolopoulos Y., "Effective Prediction of Web-User Access: A Data Mining Approach", WEBKDD 2001
22. Norguet, J., Zimányi, E., and Steinberger, R., "Semantic analysis of web site audience". ACM Symposium on Applied Computing pp. 525–529, 2006
23. Pirolli P. L., and Card S. K. (1999) Information foraging. Psychological Review. 106: pp. 643–675
24. Pirolli, P., Pitkow, J. and Rao, R. Silk from a sow's ear: Extracting usable structures from the web. ACM CHI, pp. 118–125, 1996
25. Pitkow, J. Characterizing browsing strategies in the World-Wide Web. Computer Networks and ISDN Systems, 27(6) pp. 1065–1073, 1995
26. Shardanand U., Maes P., "Social Information Filtering: Algorithms for Automating Word of Mouth", ACM CHI, pp. 210–217, 1995
27. Szalay A. S., Gray J., Thakar A. R., Kunszt P. Z., Malik T., Raddick J., Stoughton C., vandenBerg J. The SDSS SkyServer - Public Access to the Sloan Digital Sky Server Data. ACM SIGMOD pp. 570–581, 2002
28. Wang X., Abraham A., and Smith K., "Intelligent web traffic mining and analysis", Journal of Network and Computer Applications, Volume 28, Issue 2, pp. 147–165, 2005
29. White, R. W. and Morris, D., "Investigating the querying and browsing behavior of advanced search engine users", Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp. 255–262, 2007
30. Yang H., Parthasarathy S., "On the Use of Constrained Associations for Web Log Mining", Lecture Notes in Computer Science, 2703, pp. 100–118, 2003
31. Yang Q., Zhang H., Li T., "Mining Web Logs for Prediction Models in WWW Caching and Prefetching", Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 473–478, 2001
32. Zaïane O., Xin M., Han J., "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", Advances in Digital Libraries, p. 19, April 22–24, 1998
33. Zukerman, I. and Albrecht, D. W. 2001. Predictive Statistical Models for User Modeling. User Modeling and User-Adapted Interaction 11, 1–2, pp. 5–18, 2001