Multimodal Usage Visualization for Large Websites

Bibek Bhattarai, Mike Wong, Rahul Singh
Department of Computer Science, San Francisco State University, San Francisco, CA
bdb@sfsu.edu, mikewong@sfsu.edu, rsingh@cs.sfsu.edu

ABSTRACT
Large websites pose the following challenges for comprehension of user behavior: users' behaviors are complex and diverse, the web log data is very noisy, and the quantity of the web log data is of a magnitude that defies direct analysis. In this paper we present an integrated multimodal approach for usability analysis of large websites. Our research combines web content mining and web usage mining techniques in a novel integrated system which visualizes usage patterns and user goals. Furthermore, it compares web usage with website structure of content, giving a measure of design quality. Bringing usage mining and content mining together allows web designers to discover root-cause problems in the web design. The system displays multimodal information in a reflective interface. This interface provides a direct interactive visualization and query environment to discover web usage patterns. For any given usage pattern, our system uncovers the related information goal, using webpage semantic analysis and information foraging techniques. We evaluate our technique and demonstrate our system's value to improve web design and understand web usage.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Storage and Retrieval – *clustering, information filtering, query formulation.*

Keywords
Web usage visualization, information foraging, user sessions, multimodal analysis, information scent, information retrieval, visualization, mining.

1. INTRODUCTION
Web engineering today faces the enormous challenge of iteratively improving website design to facilitate use. Web designers must re-engineer websites to manage vast amount of multimedia information and make this information accessible. The success of any individual website is dependent on web users finding the information they seek (their *information goal*). Website usability is a measure of the ease by which users find their information goal. *For e-commerce sites, better web usability generates more purchase decisions, which increases revenue. For educational sites, web usability means providing relevant information to the community, thereby increasing the utility and prestige of the educational institution.*

Improving web usability means increasing the probability that users find their information goal. Web designers would like to discover ways to restructure their website to increase usability. To accomplish this, it is necessary to visualize patterns in how users are using the website (the *usage patterns*) and quantify the success for each usage pattern. Usage patterns can be thought of as a common pattern between groups of *user sessions*, which can be thought of as a series of pages that a user visits within a span of time (rigorous definition follows later).

Developing an understanding of users' usage patterns and information goals is critical in making web content accessible. The usage patterns are captured in the web log data. Web designers need a means to reorganize the log data into sensible views to detect these usage patterns. *Our research combines web content mining and web usage mining techniques in a novel integrated system which visualizes usage patterns and user goals. As a prototypical example of a large website featuring multimodal information and sustaining various types of usage patterns, we use SkyServer.*

1.1 Challenges in Analyzing Web Usage
Some of the significant challenges in analyzing web usage include:

- **Websites are designed for change.** Inactive websites suffer from data decay, or decreasing accuracy and utility of aging data; therefore active websites must continuously change their content to stay relevant. Likewise, websites must regularly restructure to accommodate growth and support new features. Constant change presents a difficulty for content mining. For example, SkyServer incorporates a revision number in their URLs; when a newer revision of the data is released, many of the pages are moved, effectively deleting the content.
- **Multimodal sites offer multiple interfaces, each having some distinct usage patterns.** Analyzing multimodal sites requires a more general definition of a usage pattern. For example, SkyServer offers several modes of user interaction: static content browsing, text-based queries (SQL), and querying by clicking on an image of the sky (receiving textual and visual feedback). To be able to capture usage patterns more accurately, a log analysis system must understand the ways a user relates with each of these modes of interactions and parse the web logs accordingly.

User Session 1:
http://skyserver.sdss.org/dr2/en/
http://skyserver.sdss.org/dr2/en/sdss/
http://skyserver.sdss.org/dr2/en/sdss/data/data.asp
http://skyserver.sdss.org/dr2/en/sdss/instruments/instruments.asp

User Session 2:
http://skyserver.sdss.org/dr2/en/
http://skyserver.sdss.org/dr2/en/sdss/
http://skyserver.sdss.org/dr2/en/sdss/data/data.asp

Figure 1. User sessions for SkyServer

- **Usage log mining only provides information about how users are consuming data, not why.** Usage log mining provides information about how users are using the website to fulfill their information goal. But it cannot provide information about what the users' information goals are. For example, consider two different users sessions for SkyServer [1] website as shown in **figure 1**. Log analysis shows that

both of the user sessions followed the similar browsing path. It also shows that user 1 visited *instruments.asp*, while user 2 left from *data.asp*. But, log analysis can not provide information about, w*hy both users followed similar path but chose different page to exit? What was their respective information goal?* Web content analysis uncovers that both users were interested SDSS data processing methodology. It also reveals *user 1* was also interested in instrument used in SDSS project to capture data. Thus, web content analysis helps us uncovering differences in information goal for the given two user sessions that followed similar browsing path.

2. PRIOR RESEARCH

This paper describes the current state of web usage mining and the features of some of the currently available tools. We then discuss how we achieve a completely different experience from currently existing tools.

Web usage mining has been a growing field [16, 9, 10] over the past decade as website developers struggle to understand their users' usage patterns (and thereby predict the users' needs). Many current products try to fill this business need and claim to have the following features: usage pattern discovery, and pattern clustering (called *user segment discovery*), and content similarity (*e.g.* product co-occurrence). Because the field is fairly mature, there is a well understood methodology which has been refined over the years. We refer to [22] and repeat the basic steps only to set the stage for discussion on the respective challenges our system overcomes.

One of the common approaches found in prior research is extraction of usage patterns from the usage log data [9, 10, 20, 22]. One of the notable challenges in pattern discovery includes clustering user requests into user sessions [16]. Discovering user sessions is known to be a difficult problem, and there are a variety of approaches [2, 14, 9], each with differing degrees of applicability for a given usage log. Other studies have attempted using machine learning and visualization techniques [5, 14] to extract usage patterns. These approaches perform usage analysis by identifying and clustering user sessions. Research works based solely on this approach do not take into account web content, and therefore cannot analyze the user goal as it relates with a usage pattern. Furthermore, after clustering the usage patterns, this approach does not suggest actionable improvements to web design.

Another common approach is web content based analysis [3, 4, 7, 20]. In this approach, researchers try to extract the information goal for a browsing pattern. They also cluster the patterns based on an extracted information goal and web content. Research works based solely on this approach do not use the information provided by the usage log. For example, they do not analyze the usage pattern in the context of the time the usage pattern occurred.

Our research philosophy radically differs from prior research. Many web systems focus on a single interpretation of what it means for a usage pattern to be interesting, since that is the algorithm that is implemented. Instead, we present a novel user interface to let the web designer decide what it means for a usage pattern to be interesting.

The proposed system begins with an *interactive visualization interface*, where the visualization interface and the query interface are integrated. We discuss the details of our interactive visualization interface later in this paper. A web designer can use this interactive visualization interface to begin a line of querying. Each query has the potential of presenting the web designer with a formulation to ask new queries. In this way, our system allows web designers to intuitively drill deeper into the underlying meaning.

By itself, usage pattern analysis is insufficient because analyzing only usage patterns does not give a clear indication of what were the users' goals. To address this issue, we correlate a usage pattern with users' goals. This correlation also helps to measure the usability of the website and identify the web design flaw root-causes.

Some of the prior research work formulates a web usage mining technique based on content mining of the website [7, 3, 4, 20]. They extract users' information goal related to a particular usage pattern. Because usage analysis based solely on web content information is incomplete, they can not provide the global picture of overall web usage. We propose a system which overcomes this limitation by combining together the information retrieved from usage mining and content mining.

3. OVERVIEW OF THE PROPOSED APPROACH

Our research focuses on bringing together the information provided by the web log data and the information provided by web content. Information such as time of use, location of the user, and what the user accesses is all available from the web logs. Information such as website structure can be directly mined from the web content, and semantic data can be inferred from content mining. Bringing these different aspects of web usage together creates a more complete experience and understanding of the usage patterns.

The proposed system starts with web usage mining by means of usage visualization. The web log information is categorized by several aspects (*e.g.* time, location, session, traffic, which pages are being accessed, *etc.*). Aspects that have an existing mental model representation (*e.g.* timelines for time, and maps for location) are displayed using those existing models as components in an integrated interface. Information is presented in the most sensible representation that preserves context. The interface is a *reflective* interface, which means that making a change in one of the components propagates the change to all other components; that is, changes *are reflected* in the other components. Using the reflective interface of the system, web designers are able to analyze the usage log and discover usage patterns from relationships between different aspects. This includes spatial and temporal characteristics related to the usage pattern. This system lets web designers perform further mining of a usage patterns to uncover information goal.

The web designer can then mine the content of the website in the context of the usage pattern. If the results of content mining are logically consistent with the web designer's knowledge of the domain and website structure, then the web designer can use all of this information to draw interesting conclusions such as the case study experiment described later. If the results are logically inconsistent, then the web designer can dismiss the apparent usage pattern as an inliers. This is just one of the advantages of combining web usage and web content mining.

First, the usage log data is preprocessed to remove erroneous data and noise. Then usage patterns are uncovered from the usage log data. Finally, with user sessions and other patterns defined, the

usage patterns are analyzed. The patterns can be analyzed in a variety of ways, including searching for clues for web system improvement, website improvement, capturing a larger revenue stream, and user characterization.

We perform web content mining based on Natural Language Processing (NLP) and information foraging techniques. First, we analyze web content using a common NLP technique known as Term Frequency Inverse Document Frequency (TFIDF) [23]. This technique predicts the importance of a term in a document depending on how commonly the term occurs in a given document collection.

Then we use a technique based on Information Foraging theory [19] to predict the user information goal and the user flow in a website. This technique is based on the known fact that users predict the content of a distal page based on the information hint provided by the link pointing to the page [19].

Using Breadth First Search (BFS) we compute the shortest distance (optimal) path between a user's start page and final goal page. Then using directed graph we construct graphical representation of the user's path, optimal path and the user flow prediction related to the usage pattern.

Graphical representation of the web structure and usage data can help web designers to perform better visual analysis of web usability. However, structural and usage graphs of large websites are overwhelmingly complex and not useful for visualization and understanding. In this system we overcome the problem by presenting a context-based subset of the web structure and usage graph that are related to the usage pattern being analyzed.

3.1 Benefits of Integrating Web Usage Visualization and Web Content Mining

The approach taken in this paper combines two separate technical perspectives: web usage analysis using multimodal visualization and content mining with structural and user flow visualization. By uniting these two components, we address the challenges listed above (Section 1.1) the following ways:

- **Multimodal Visualization is a general technique.** Visualization is completely data-dependent; it is unaffected by rapidly changing data and does not make any presumptions beyond the data schema. Outliers in the data are self-identifying. Given that over 92% of surveyed web servers [26] have a common subset of log information, visualizing this subset is general to most websites. The advantage of an interactive, unified visualization/query UI is that a user can drill-down and recursively query non-outliers to differentiate good data from inliers and recognize patterns. Visualization systems which offer multiple modes of information (such as timelines to display temporal information, maps to display spatial information, and charts to discover relationships between numeric and enumerated data) offer a more complete view of the web log information than a simple listing. Because these modes are all associated with the web log data, allowing a user to manipulate any of these dimensions creates a querying environment that leads to deep understanding.
- **Visualization and content mining are both enhanced by domain knowledge.** Research shows that people can identify complex relationships in patterns in a few seconds [27]. Some prior works on usage pattern mining take this for granted in choosing what patterns are "interesting" [8, 13]. Therefore, relying on visualization techniques is at least as good as similar judgment-based works. *Most importantly, identifying important usage patterns requires domain knowledge and knowledge of the web site design, especially for multimodal systems.* Therefore a system which promotes free pattern recognition through visualization provides a more general solution. Finally, a web designer with domain knowledge can use the results of content mining to validate or dismiss an apparent visualized usage pattern. That is, Integrating web usage mining and content mining provides the complete picture. Web usage mining tells web designers how web users are using the website to fulfill their information goal. Web content mining tells us what users are looking for, that is their information goal related to a particular usage pattern.

4. SKY SERVER AS A MODEL LARGE WEBSITE

For experimental evaluation of our system we use real usage data collected from SkyServer [1]. The goal of this website is to provide web access to Sloan Digital Sky Survey (SDSS) [26] data through standard web browsers. The website provides various types of interaction between the user and the data [25],

- 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

The targeted audience for this website ranges from a kids learning astronomy at their school to research scientists and astronomer. A well formulated design is must to provide easier and optimal access to the SDSS data for the wide range of users.

SkyServer is a *very* large website, offering views and data for over 80 million astronomical phenomena, totaling over one-and-a-half terabytes. Our copy of the usage log data is approximately 35 gigabytes and only spans from May of 2003 to October of 2004.

One of the challenges in dealing with freely accessible large websites is that many of the usage patterns are too complex to cleanly distinguish one from another.

Another challenge is that the signal-to-noise ratio is very high. That is to say that there are very large numbers of information-poor web requests (*e.g.* erroneous web requests, web service error conditions, button images, navigational images, logos, cascading style sheets) relative to informative requests (*e.g.* successful web service requests, web pages and images of astronomical phenomena).

Simple aggregate analysis of the web logs reveals:

- less than 0.4% of all web requests are informative requests
- approximately 38.3% are made by web robots
- approximately 17.4% of these web robots are using a single web service
- Of the requests made by humans, over 73.3% of the user sessions don't progress past three pages