ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

User Data model detection, and Pre-processing Pattern analysis

Vijetha Rudra
Assistant Professor
Department of IT
Teegala Krishna Reddy Engineering College
Medbowli, Saroornagar, Hyderabad.

Pavani Somireddy
Assistant Professor
Department of IT
Teegala Krishna Reddy Engineering College
Medbowli, Saroornagar, Hyderabad.

Abstract
In the real world, lot of users attracted towards online shopping, so lots of transactions are going on in the websites. A weblog contains series of entries updating frequently by the user while accessing the website. Based on the user interest, it can be classified as related and unrelated data. The related data can be considered as success response, but the unrelated data can be considered as failure response. It analyze the pattern of user navigation while browsing, for that web usage mining must be analyzed. The steps consisting for the process of web usage mining are data collection, pattern discovery, and Pre-processing and pattern analysis.

Keyword: Cognitive user model, sessionization, software tool, test oracle, usability, web server log.

I. INTRODUCTION

Online shopping (sometimes known as electronic retail or e-shopping) is a form of electronic commerce which allows consumers to buy goods or services from a seller over the internet using a web browser. Alternative names are: e-web-store-shop, e-store, internet shop, online store, and virtual store. Mobile commerce describes about online retailers in mobile optimized online site or app. online customers mostly use a credit card in order to make payments. Few online shops will not accept international credit cards. Few of them requires both the purchaser‘s billing and shipping address to be in the same country. Other online shopping allowing customers to any country to send gifts anywhere.

Customers searches the product of interest by visiting the website of the retailer directly or by searching among alternative vendors by a shopping search engine. Once a particular product was found on the website, most of the online retailers will chose shopping cart software to allow the consumer to accumulate multiple items and adjust quantities like filling a physical shopping cart or basket in a conventional store. Some stores allow consumers to sign up for a permanent account usage. Less sophisticated stores may reply on consumers to phone or e-mail their orders.

A cognitive model is an approximation to animal cognitive processes (predominantly human) for the Purposes of comprehension and prediction. Cognitive models can be developed within or without a cognitive architecture, though the two are not always easily distinguishable. Sessionization refers to the capture of all click stream activity within the timeframe of a single visitor‘s Website session. A programming tool or software development tool is a computer program that software developers use to create, debug,

maintain, or otherwise support other programs and applications. Software testers and software engineers can use an oracle as a mechanism for determining whether a test has passed or failed. In Software engineering, usability is the degree to which a software can be used by specified consumers to achieve quantified objectives with effectiveness, efficiency, and satisfaction in a quantified context of use. A server log is a log file (or several files) automatically created and maintained by a server consisting of a list of activities it performed. A typical example is a web server log which maintains a history of page requests.

Online customers must access the internet and a valid method of payment to complete the transactions. Generally higher levels of education and personal income correspond to more favorable perceptions of online shopping. Increased exposure to technology also increase the probability of developing favorable attitudes towards new online purchasing. Customers

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

are attracted to online shopping not only because of high levels of convenience, but also because of border selections, competitive pricing and greater access to information. Business organizations seeks to offer online shopping not only because it is lower cost compared to bricks and mortar stores, but also it offers a worldwide market, increases customer value, and builds sustainable capabilities.

Designers of online shop are carrying the effects of information load. Information load is the product of spatial and temporal arrangements in web store. Comparing with conventional retail shopping, it provides additional product information such as comparative products and services, with various alternatives and attributes. Two major dimensions of information load are complexity and novelty. Complexity means number of different elements or features of a site, the result of increased information diversity. Novelty means the unexpected, suppressed, new or unfamiliar aspects of the site.

A successful web store is not a good looking website with dynamic technical features, listed in many search engines. Business often attempts to adopt online shopping techniques without understanding them without a sound business model. Business produce web stores meets organizations culture and brand name without satisfying consumer expectations. User-centered design is critical and complex. Understanding the customer needs is essential. Living in company‘s promises gives customers a reason to come back and meeting their expectations. It is important that the website communicate how much the company values to customers.

Customer needs and expectations are not equal for all customers. Age, gender, experience and culture are all important factors for online shopping. Users with online experience focus more on the variables that directly influence the task, while other users focus on understanding the information. To increase online purchases, businesses will be using significant time and money to define, design, develop, test, implement and maintain the web store. It is to eliminate mistakes and more appealing to online shoppers, many web shop designers study research on consumer expectations.

II. RELATED WORK

Matthias Rauterberg suggested that ―To support the human factors engineer in designing a good user interface, a method has been developed to analyze the empirical Oataófttre interactive user behavior traced in a finite discrete state space. The sequence of actions produced by user contain valuable information about the mental model of their, the individual problem solution strategies for given task and the hierarchical structure of the task--sub tasks relationships. The presented method, AMME, can analyze the action sequences and automatically generate (l) a net description of the task dependent model of the user, (2) a state transition matrix, and (3) various quantitative measures of the user's task solving process. The behavioral complexity of task-solving processes carried out by novices has been found to be significant than the complexity of task-solving processes carried out by experts‖.

Michigan and Ann Arbor discovered that ―Weblog are analyses for the way of IT administrators to ensure adequate bandwidth and server capacity on organizational web sites. In the past five years, Log file analysis has advanced with companies; now mining files are fine-grained detail about visitor profiles and buying activity.

Organizations are now seeks to use log files to learn about the usability of web sites—that is, successful visitor meets their specific information or transaction goals. Log file data can offer valuable insight into web site usage. It reflecting actual usage in natural working condition compared to artificial settings of a usability lab. It represents activity of many users, over potentially long period of time visited to a limited number of users for an hour or two each‖.

Linda Tauscher and Saul Greenberg discovered that We report on users‘ revisitation patterns to World Wide Web pages, and empirical foundation for the design of history mechanisms in web browsers. Through history, it returns to visited page by reducing the cognitive and physical overhead required to navigate for scratch. People tends to revisit pages for accessing web pages frequently, browse in very small clusters of pages, and generates short sequences of URL paths. We compared different history mechanisms, and found the stack-based prediction method prevalent in commercial browsers is inferior to simpler approach of recently visited URLs with duplicates removed.

Fabio Paterno and TonioCarta discovered that Usability evaluation of Web sites is time consuming task performed manually. It presents a tool that supports remote usability evaluation of Web sites. It supports client-side data's on user interactions and JavaScript events. Moreover, the definition of custom events giving evaluates the flexibility to add specific

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

events to detected and considered for evaluation. The tool supports the Web site by exploiting a proxy-based architecture and enables the evaluator to perform actual user behavior and an optimal sequence of actions‖.

Jeff Tian, Sunita Rudraraju and Zhao Li discovered that Based on the characteristics of web applications and web environment, we classified web problems and focus on the subset of source content problems. Using information about web accesses, we derived various measurements of web site workload at different levels of granularity and various perspectives.

The workload measures combined with failure information extracted from recorded errors. It evaluates the operational reliability for source contents at potential for reliability improvement. The given results demonstrates the viability and effectiveness of our approach.

Marcus Heinath, Jeronimo Dzaack, Andre Wiesner discovered that ―Usability of complex dynamic human computer interfaces can be evaluated by cognitive modeling for their underlying structures. Even though the prediction of human behavior can check the detective errors in interaction design and cognitive demands of the future.

HTAmap, a high-level framework for cognitive modeling, it aims to reduce the modeling effort. Within HTAmap the cognitive models transformed into pattern-oriented model. This paper describes both concepts and first implementations. Both tools was shown using an example of process control‖.

III. SYSTEM ARCHITECTURE

In system architecture, the user will browse the website for online shopping. After accessing the website, they sends request to Web server and web server sends the response to web browser (HTTP request and HTTP response).The hypertext transfer protocol (HTTP), it is an application protocol for distributed, collaborative, hypermedia information systems (Fig 1.1).

Hypertext is a structured text that uses logical links between nodes containing text. The series of entries between web browser and web server is stored in weblogs. Weblog is a raw log using this pre-processing technique will be activated.

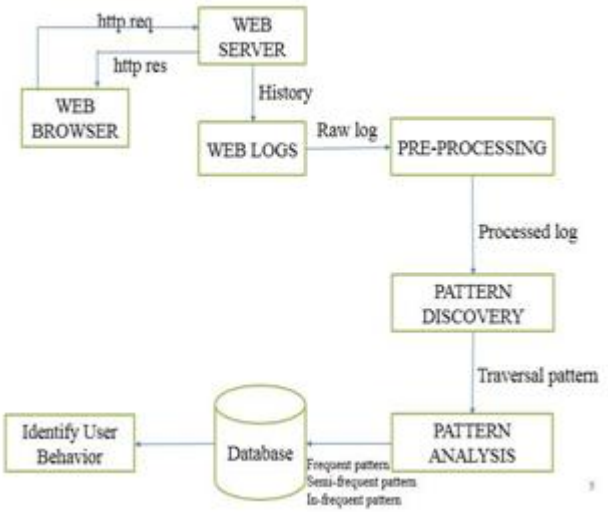

Fig 1.1

Online shopping is a form of electronic commerce it allows consumers to buy goods or services from a seller over internet using web browser. Some online shops will not accept international credit cards. Few of them require both the purchaser billing and shipping address to be in the same country. Other online shops allow customer‘s from any country to send gifts anywhere. Online customers must access the internet and a valid method of payment to complete the transactions. Generally higher levels of education and personal income corresponds to more favorable perception of online shopping (Fig 1.1). Increased exposure to technology also increase the probability of developing favorable attitudes towards new online purchasing.

The new method to identify navigation related usability problems by comparing Web usage patterns extracted from server logs against anticipated usage represented in some cognitive user models. It includes three major modules: They are: Usage Pattern Extraction, Internet Usage for Information Provisioning (IUIP) Modeling, and Usability Problem Identification. First, it extracts actual navigation paths from server logs and discover patterns for typical events. In parallel, it will construct IUIP models for the same events. IUIP models are cognition of user behavior and it can represent anticipated paths for specific user-oriented tasks. The result checking employees the mechanism of test oracle. An oracle is generally used to test that passed or failed. Here, we use IUIP models as the oracle to identify the usability issues related to the users‘ actual navigation paths by analyzing the deviations between the two.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

IV. DESCRIPTION METHOD

Implementation is the theoretical design is turned out into a working system .Each entry in a log contains the IP address of the originating host, time stamp, requested Web page, referrer, user agent and other data. Typically, the raw data need to be Pre-processed and converted into user sessions and transactions to extract usage patterns.

A. Log generation

In log generation module, the user searches the query and that hits the server as http request and server responds back to the user as http response. Those http request and response is considered as web logs. The prepared raw weblog dataset contains null values and unwanted datasets. Pre-processing technique is used to remove the null values in web search log dataset (Fig 1.2). After the preprocessing process is completed, it will be moved on to Pattern generation module. In pattern generation, single customer details will be covered. And then it can be classified into three categories: Frequent, semi-frequent and in-frequent. Clearing denotes all activities from the time a commitment is made for a transaction until it is settled. Clearing of payments is necessary to turn the promise of payment into actual movement of money from one bank to another.

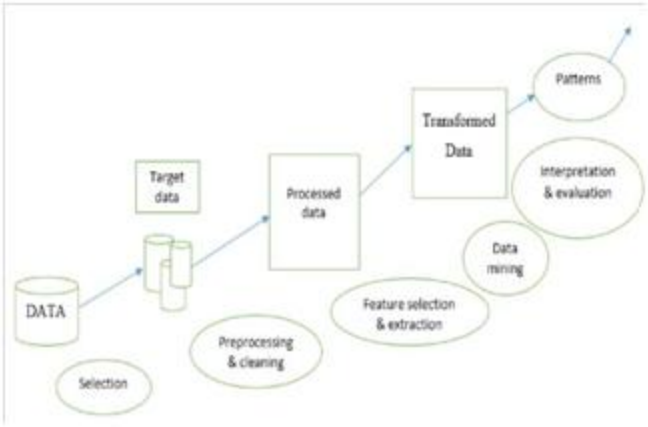

Fig 1.2

In this phase, the application of web log data may need to be cleaned from entries such as pages that returned an error or graphic file accesses. In such cases few information might be useful, but in other such data should be eliminated from log file.

Therefore, crawler activity can be filtered out, because such entries not provide information about the site‘s usability. Another problem is too met with caching. Cached pages are not recorded in the web log. Caching is heavily depending on the client-side technologies. Therefore cannot be deal with easily.

B. Weblogs

A weblog that consists of a series of entries updated on frequently with existing information. The information can be written by the site owner, or other sources, or contributed by users. Generally, weblogs are usually of topical interest, and, in general, it can be thought of as developing commentaries, individual or collective on their particular themes (Fig 1.3). A weblog consists of the recorded ideas of an individual (a sort of diary) or be a complex collaboration .Since there are a number of variation on this idea can be easily invented, the meaning of this term is to gather additional connotations with time.

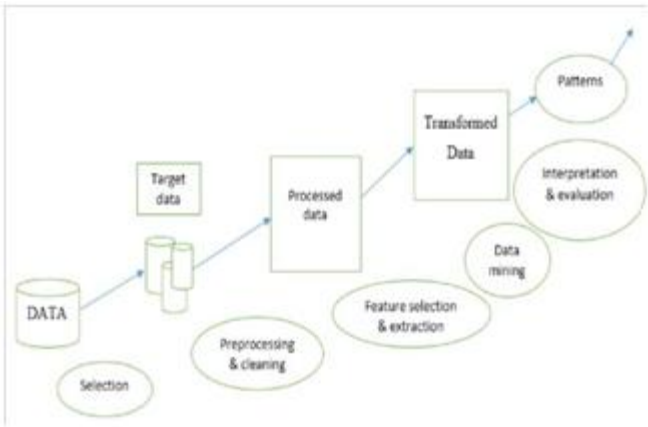

Fig 1.3

A weblog is an informational site published on the World Wide Web consisting of discrete entries ("posts") typically displayed in the most recent post appears first. Weblogs were usually work in single individual group. Recently, "multi-author blogs" have introduced with posts written by large numbers of authors and professionally edited. It is from newspapers, other media outlets, universities, think tanks, advocacy groups and similar institutions account for increasing quantity of blog traffic.

C. Logs Preprocessing

The data preparation and preprocessing includes the following domain-dependent tasks.

1) Data cleaning: This task is usually site-specific and removing extraneous references to style files, graphics, or sound files that may not be important for the purpose of analysis.

2) User identification: The remaining entries are grouped by individual users. Because no user authentication and cookie information is available in most server logs, we used the combination of IP, user agent, and referrer fields to identify unique users.

3) User session identification: The activity

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

record of each user is segmented into sessions, with each representing a single visit to a site. Without additional authentication information from users and without the mechanisms such as embedded session IDs, one must rely on heuristics for session identification. For example, set an elapse time of 15 minutes between two successive page accesses as a threshold to partition a user activity record into different sessions (Fig 1.4).

4) Path completion: Client or proxy side caching can often result in missing access references to some pages that have been cached. These missing references can often be heuristically inferred from the knowledge of site topology and referrer information, along with temporal information from server logs.

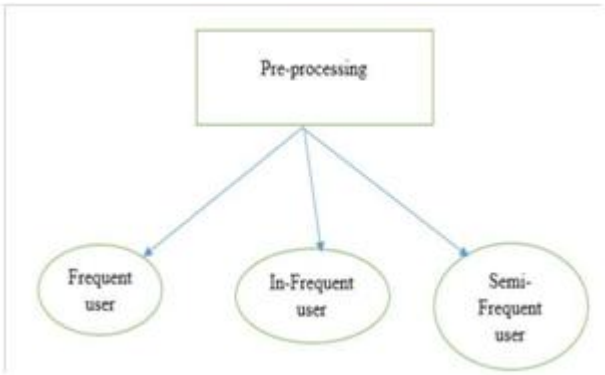

Fig 1.4

D. Pattern Generation Module

In Pattern Generation, analyzing the web log files through web usage mining is very important to discover the similar behavior users of particular website. Our paper discusses how to find useful knowledge from web log file using some data mining technique like Association rule mining and clustering. First we pre-process the web log file then apply association rule mining and P.G.Vedaprakash*etal*, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 90-99 clustering algorithm on web log file to discover usage pattern and same behavioral users. Classification consists of predicting a certain outcome. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and outcome called goal or prediction attribute (Fig 1.5). The algorithm tries to discover the relationship between attributes would make it possible to predict the outcome.

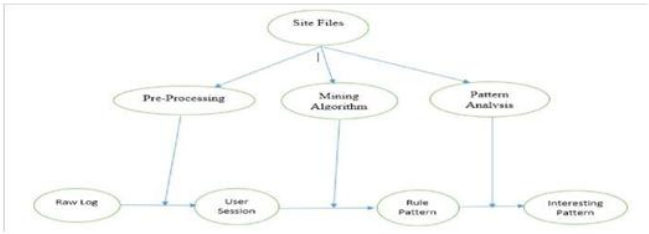

Fig 1.5

Classification consists of predicting a certain outcome based on given input. In order to predict the outcome, the algorithm processes a training set with attributes and the respective outcome, also called goal or prediction attribute. The algorithm tries to discover the relationship between the attributes makes it possible to predict the outcome. Classification consists of Supervised and Unsupervised. Supervised Classification- Set of possible classes is known in advance. Unsupervised Classification- Set of possible classes is not known. It is also called clustering.

E. Pattern Analysis

Web usage mining has many advantages which makes the technology attractive to corporations including the government agencies. This technology has personalized marketing, which eventually results in higher trade volumes. Government agencies are classified its threats and fight against terrorism. The predicting capability of mining applications can identifying criminal activities. The companies can establish its own customer relationship by giving them exactly what they need.

The companies can find, attract and retain customers depends on saving the production costs by utilizing the acquired insight of customer's requirement. They can increases the profit by targeting price based on the profiles created. They can even finds the customer who might find the default to a competitor the company will try to retain the customer by providing promotional offers to the specific customers, thus reducing the risk of losing a customer or customers.

We adapted the tire algorithm to construct a tree structure that also captures user visit frequencies, which is called a trail tree in our work. In a trail tree, a complete path from the root to a leaf node is called a trail. Each node corresponds to the occurrence of a specific page in a transaction. It is annotated with the number of users having reached the node across the same trail prefix (Fig 1.6).

Clustering is classified into three types:

1) Structure-based Clustering - Clustering result

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

based on vertex connectivity. They are closely connected; e.g. half work on XML and the other half work on Skyline in one of the cluster.

2) Attribute-based Clustering - Clustering results based on attribute similarity. Within clusters work on the same topics, however, the relationship may be lost due to the partitioning so that author are quite isolated in one of the clusters.

3) Structure/Attribute Clustering - Clustering results based on both structure and attribute similarities. Within one cluster are closely connected, meanwhile, they are homogeneous on research topics.

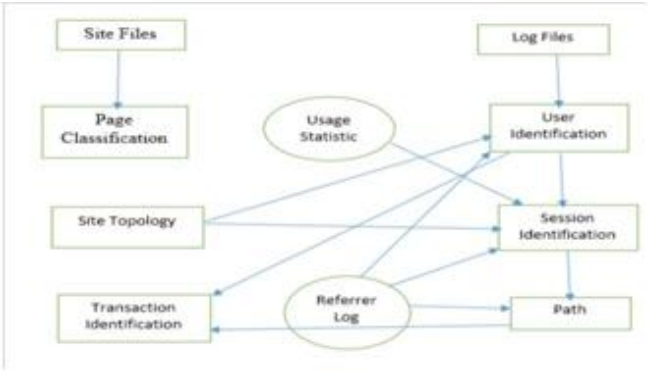

Fig 1.6

F. Identify the user Behavior

In graph modeling, the Web usage mining and statistical analysis are the two ways to evaluate usage of Web site. Web usage mining and statistical analysis considers client side data. In other words, it combines graph based Web usage mining and browsing time analysis while taking client side data into account. It helps us to reconstruct user session exactly by based on these data, we find Web usage patterns with more accuracy.

Some non-digital products have been more successful than others for online storage. Profitable items have a high value-to-weight ratio, they involve embarrassing purchases, and typically go to people in remote locations, and they may have shut -ins as their typical purchasers. Items can fit in a standard mailbox—such as music CD's, DVDs and books—are particularly suitable for a virtual marketer (Fig 1.7).

Based on the ranking process classify the weblog data-set into frequent and infrequent item set. Here, frequent item set indicates the user navigation. Then, forward the frequent item set and navigation to the respected user. This the procedure where information stored in web server logs is processed by data mining techniques in order to,

1) Extract information and discover interesting usage patterns.
2) Cluster the users into groups according to their navigation behavior.
3) Discover potential correlation between web pages and user groups.

The process of extracting information concerns the browsing behavior of users can be regarded as part of the user profiling process. It is evident that the user profiling and web usage modules overlap. In many cases information provides many web site is not physically stored in the web site‘s server. In case of a web portal, users are interested in information from various web sources. The web site editors are the contents of interest should consequently be classified into two categories. Searching and relevance ranking techniques are employed in the process of acquisition of relevant information and in the publishing of the appropriate data to each group of users.

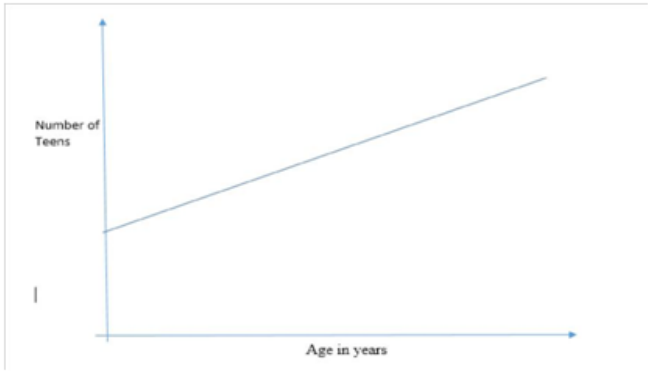

Fig 1.7

Our IUIP models are based on the cognitive models surveyed in the Adaptive Control of Thought—Rational (ACT-R) model. Due to the complexity of ACT-R model development and the low-level rule based programming language and constructed our own cognitive architecture and supporting tool based on the ideas from ACT-R.

V. CONCLUSION AND FUTURE WORK

The extraction of related information from web browser is identified the frequent and semi-frequent customers for online-shopping. Using this type of information, users will save time management and quality products. In future, Web usage mining is the base for navigation pattern mining and approach of clustering is used to perform that Mining, usage mining deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications and also to improve the customer relationship management. I would like to conduct more comprehensive experiments to further verify my approach and improve upon it.

REFERENCES

1. A. McDonald and R. Welland, ―Web engineering in practice, in Proc. 10th Int. World Wide Web Conf., May 2001, pp. 21–30.

2. M. Rauterberg, ―AMME: An automatic mental model evaluation to analyse user behaviour traced in a finite, discrete state space, Ergonomics, vol. 36, no. 11, pp. 1369–1380, 1993.

3. P. M. Sanderson and C. Fisher, ―Exploratory sequential data analysis: Foundations, Human–Comput. Interaction, vol. 9, nos. 3/4, pp. 251–317, 1994.

4. L. Tauscher and S. Greenberg, ―Revisitation patterns in World Wide Web navigation, in Proc. ACM SIGCHI Conf. Human Factors Comput.Syst., New York, NY, USA, 1997, pp. 399– 406.

5. G. Christou, F. E. Ritter, and R. J. Jacob, ―CODEIN—A new notation for GOMS to handle evaluations of reality-based interaction style interfaces, Int. J. Human-Comput. Interaction, vol. 28, no. 3, pp. 189–201, 2012.

6. Tec-Ed, ―Assessing web site usability from server log files, White Paper, Tec-Ed, 1999.

7. T. Arce, P. E. Roman,´ J. D. Velasquez,´ and V. Parada, ―Identifying web sessions with simulated annealing, Expert Syst.
Appl., vol. 41, no. 4 , pp. 1593–1600, 2014.

8. M. Heinath, J. Dzaack, and A. Wiesner, ―Simplifying the development and the analysis of cognitive models,‖ in Proc.
Eur. Cognitive Sci. Conf., Delphi, Greece, 2007, pp. 446–451.

9. F. E. Ritter, A. R. Freed, and O. L. Haskett, ―Discovering user information needs: The case of university department web sites, ACM Interactions, vol. 12, no. 5, pp. 19–27, 2005.

10. J. S. Valacich, D. V. Parboteeah, and J. D. Wells, ―The online consumer‘s hierarchy of needs,‖ Comm. ACM, vol. 50, no. 9, pp. 84–90, Sep. 2007.