RED: A Model To Analyze Web Navigation Patterns

Eduardo Marques
Universidade Federal Fluminense
Rua Passo da Pátria 156/456
RJ, Brasil CEP 24210-240
+5521 26229361
emarques@urbi.com.br

Ana Cristina Garcia
Universidade Federal Fluminense
Rua Passo da Pátria 156/456
RJ, Brasil CEP 24210-240
+5521 26295675
bicharra@dcc.ic.uff.br

Inhaúma Ferraz
Universidade Federal Fluminense
Rua Passo da Pátria 156/456
RJ, Brasil CEP 24210-240
+5521 26295675
ferraz@dcc.ic.uff.br

ABSTRACT
Internet has become a popular medium to get and provide information and services. Every user is a potential client, and every site a potential competitor. Surfing the Web produces a trace that hides important information concerning business opportunities and site usability issues. This paper discusses the use of an intelligent agent to interpret Web navigation history and act opportunistically to take advantage of business opportunities. We argue that such an agent should work off-line discovering relevant navigation patterns and work on-line acting upon this information to send messages or change a Web page's content, whenever a relevant navigation sequence is recognized.

Keywords
Web navigational patterns, pattern recognition.

INTRODUCTION
With the growth of the Web, new types of systems become available via an Internet connection [1], what has increased the number of relevant factors when evaluating the usability of interaction.

Users' navigation registers – data logs – must be captured and serve as source of analysis to help designer evaluate the site usability and define the user's profile. Thus, the evaluation must generate quality results, suggest design solutions to the problems found, with a timely answer, and, as a subproduct, offer a customized navigation.

Our study shows a model (RED) for Web site evaluation based on navigation patterns recognition. Agent's technology and parsing algorithms are RED's key-components. Our method analyzes and diagnoses user's navigation when using a Web service to determine relevant navigation sequences.

WEB NAVIGATION SEQUENCES
Navigating through the Web is to visit site pages in a structured order (moving from page to page according to the site internal links) or in a random order (taking an unsystematic strategy such as using the *Back* button on a browser, entering a new URL address, etc.). All of these Web browsing mechanisms can be purposeful or not. Users' behavior is related to their intention and to their ability to find what they are looking for. Combining these two factors, four types of Web navigation can be outlined: Direct search access (purposeful navigation), Direct access (almost no navigation), Content-related access (navigation inside a web page), and Random access (navigation with no purpose).

Each visit to a site is recorded as an anonymous contact in a log file stored in the site server. The log can be interpreted as the user-web site's interaction. In addition to identifying users, interpreting a web site log may help diagnose the site usability, and better understand users.

Agent technology offers a way to track users even outside the scope of a web site. However, there are some legal issues involved when applying this technology, such as guaranteeing users' privacy. Therefore, we will remain within the scope of a web site's log.

A navigation trace may form patterns that can show web page designers how to improve their site attractiveness and usability. A pattern is a rule that relates a problem, a solution and a context. In the Web, a pattern can be a sequence of pages that allows users to reach a specific set of content to satisfy their needs. The problem is kept in the user's mind, the solution is within the navigation, and the context is the Web. There are many different ways to reach a desired content, some preferable over others. There are ways of optimizing sites access that have already taken advantage of observed navigation patterns, such as:

- **Set-based navigation**: information reorganized by clustering related pages by topic.
- **News**: information classified into new and old.
- **Bookmark/Favorites**: users can create their own space of subject interests.
- **Shopping basket**: function related to e-commerce.
- **Advanced search mechanisms**: users restrict their search by refining the search space or specifying keywords.

THE WEB SITE EVALUATION PROBLEM
Evaluation can be divided into two parts: the gathering of the data to be evaluated and its evaluation. The advantage of this division is that data is gathered from a real usage

environment and can be evaluated in a specialized laboratory. There are a number of ways to do these two remote evaluation steps. Among them we can point out [2]:

Commercial usability services – system's developers send documents, examples, and/or prototypes to hired remote evaluators.

Remote questionnaire or Survey – users are showed a questionnaire in which the system collects their system and interface preferences.

Collaborative remote evaluation – system's user and usability evaluator communicate through the network.

Video-conferencing-supported evaluation – user remains connected to evaluator through a video-conferencing program that broadcasts the user's screen dumps.

Instrumented or automated data collection for remote evaluation – system captures user's actions and sends them to evaluator.

User reported critical incident method – user is trained to identify critical incidents and send evaluator a report with the information.

When considering gathering data to evaluate a system's usability, it is natural to believe that the more data gathered the better the evaluation. However, there are a number of problems that makes data gathering on the Web non-trivial. The information about Web navigation is distributed through the network. Some is on the visitor's computer, and the rest is at the various sites the user is visiting. Therefore, gathering this data without loss, inconsistencies, or redundancies is a complex task.

Another problem occurs when we begin to think about information security. In order to protect the user's privacy, some personal data may not be revealed for ethical reasons, such as the content of e-mails. There is also data that cannot be gathered such as passwords, credit card numbers, etc. In order to protect these types of data, there are mechanisms that make it impossible to them. This brings us to a technological problem, since there are limits to the information that can be gathered. This limitation would not be an issue if it only took place with these types of data. However, there are cases where it is not possible to distinguish this type of information from that which would help in the evaluation. In this case, the tools themselves prevent access to all the types of information, making it more difficult to gather navigation data.

In [3] it is presented a tool (WBI) to collect the navigation data of a user in the Web, capturing the entire exchange of information between these two means of usability evaluation, without access restrictions to the information. Moreover, it contains a low transparency to the user. Still, these tools present a few problems: (a) all information necessarily passes through an intermediary, slowing navigation; moreover, depending on the quality of the connection, this can become a problem; (b) all information required for evaluation is captured with the user's personal information; but to guarantee the data's security (even if not kept or used) generates doubts; (c) the information ends up becoming homogenous due to the lack of contextualization of the actions, as there is no distinction between the type, form, or use of each action.

On this point, one can apply the heuristic evaluation; i.e., a set of human evaluators would examine the data collected during the interface's usage, and would judge it based on principles of usability, which in truth are usage patterns already known to the system. A search for usage patterns, however, would be done manually, which does not guarantee an acceptable evaluation time. It is necessary, though, to have an automatic pattern recognition method.

Once we have collected all the information on the user's navigation, we will also have the entire sequence followed during the visit to the site pages. Besides, each of the steps in the sequence has a context determined by the known parameters, such as the page's name, the values of the variables, the type of page, etc. Some sequences contain a special meaning, indicating the repetition of a sequence tends to mean the same intention for all. This meaning can be considered a navigation pattern and help the designer define navigation patterns found at a specific site.

Based on the recognized navigation patterns of users at a site, we will formulate hypotheses on their intentions and their behavior when faced with the layout of the information available at the site pages. By knowing a user's purpose in a site, we also know their state (a potential buyer, a user disappointed with a site service, etc.). It becomes, therefore, possible to personalize the navigation, adapting its content and/or the user's navigation to their needs, such as the modeling of personalized Web applications [4].

In [5] one can see a great number of Web personalization tools such as Groups Lens and LikeMind**,** and Personalization Server and FrontMind.

RED MODEL – WEB SYSTEM DATA LOGS' ANALYZER

RED[1] (Navigation Strategies Recognizer) is a model that describes a method to recognize decision strategies that a site's customer used while navigating. From this recognition, based on the user's navigation history at the site, it is possible to evaluate the use of a site and suggest actions to ease interaction.

The RED model in Figure 1 describes the behavior of four special agents: Interaction Analyzer Agent, Pattern Actuator Agent, Pattern Generator Agent, and Pattern Configuration Interface Agent, that work together to make sense of users' Web page navigation. All the agents use the server log data to act defining or finding patterns, a sequence of visited pages that may be often repeated. The pattern may or may not contain noise.

[1] Portuguese acronym

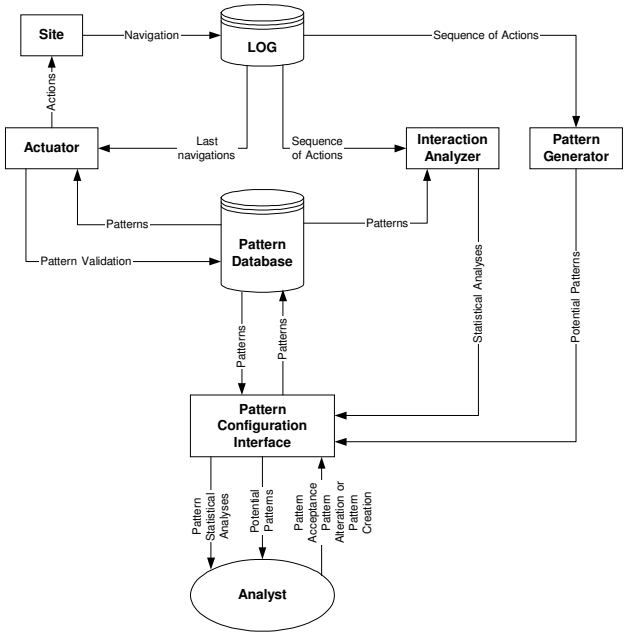

Figure 1– Model to automatically evaluate usability

Interaction Analyzer Agent: it looks in the navigation log for action sequences that are not made of known patterns, but that repeat themselves in the log. It analyzes those sequences and creates a report to be sent to the Pattern Configuration Interface, which, along with the human analyst, determines whether a pattern exists and whether it must be configured.

Pattern Actuator Agent: while the user is visiting the site pages, among the last sequences, the Pattern Actuator will try to anticipate a pattern. When a pattern is recognized, it presents an action related to the pattern in its base of actions, and validates the pattern in the pattern database.

Pattern Generator Agent: it looks in the Web log data searching for potential sequences (with or without noise) that are relevant. These sequences are then sent to the Pattern Configuration Interface, which, along with the human analyst, determines whether a pattern exists and whether it must be configured.

Pattern Configuration Interface: it allows the analyst to evaluate the patterns, not necessarily related to a specific log data, and decide whether he will include, or alter or delete existent patterns.

Pattern Database: it is a database containing pattern descriptions, its associated meaning, an application context describing pre-requisites for applying a pattern, a trust factor and an action to be taken when a pattern is recognized during a surfer's navigation in the site.

The log is the most important item within the RED model. It is where one can find everything the user did in the system. For this reason, defining its format, content, and creation becomes weighty factors of the quality of the answer reached. A log containing little information may make it impossible to detect patterns, or worse, may cause incorrect or incomplete patterns to be detected. On the other hand, a log containing too much information may make pattern detection difficult, or even impossible. As we are dealing with a system being used through the Web, information on the user's interaction will not be considered because, despite being able to obtain some of this information, the percentage of this type of information is so small that it becomes preferable to consider that we only have navigation information.

CASE STUDY

An implementation of the RED model has demonstrated the model was feasible. We will describe a prototype of the RED model for the RiskControl site [6], which offers a service for risk calculation of financial investments, the possible actions within the site environment, the patterns considered in the pattern database and the actions in the action database.

In general, the service sites have tools available to execute a given task. For a user to execute such a task, he does not need to navigate among several sites, only within the site with the desired service. Let us see the Investment Analysis Service. RED model was attached to RiskControl (Figure 2), which allows the user: create investments portfolios; calculate the risk of their portfolios; and simulate market situations.

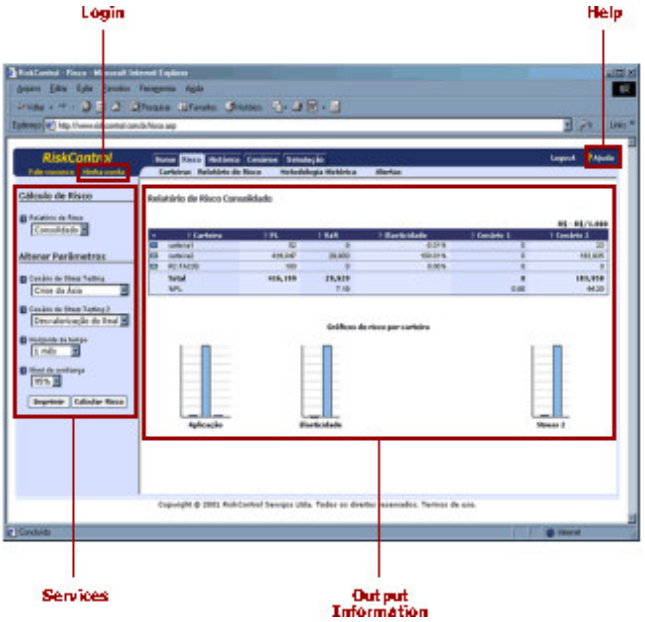

Figure 2 – Screenshot of RiskControl site

Thus, we can define as alphabet in these types of site the possible actions a user can execute when they are visiting a site (Table 1).

Table 1 - Actions in a service site

Type	Description	Related Element
INFO	Read about a subject	Subject
SERV	Use a service	Service
HELP	Request help	Subject

In the pattern database we will include the sequences of actions that have a known meaning (Table 2). These patterns make up the dictionary of the service sites.

Table 2 - Patterns in a service site

Name	Description
same subject	Read about a same subject several times.
several subjects	Read about several subjects.
same service	Use a service several times.
several services	Use several services.
same help	Request help several times on a same subject.
several helps	Request help on several subjects.

The log used by the test contains 9508 actions executed by the users in 2322 sessions from February 2001 to June 2001. Each visit to the site by a user was considered as a session formed by a sequence of actions. After the use of the Statistics Generator on the aforementioned sequences, 1689 patterns were uncovered.

As can be seen in Figure 3, most of the site use was to utilize its available services (67.92%), which is to be expected since that is the site primary function. Almost half the actions were to use a service (41.77%) and a few others were to explore the same service more than once (12.49%).

Figure 3 – Service patterns distribution

These data show that users are satisfied with the results found with the services, but the services are not being exploited by the users, either because said possibility does not exist or because they did not know how to.

The help patterns (Figure 4) were used few times (1.95%), with no hits on a number of subjects. This shows it is easy to learn how to use the site.

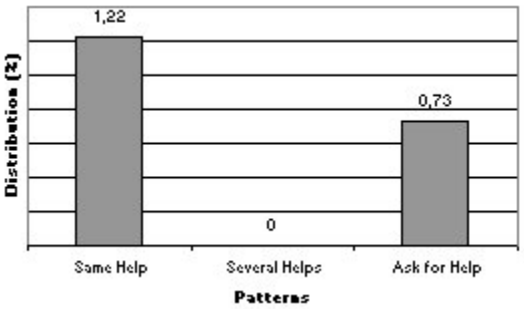

Figure 4 – Help patterns distribution

CONCLUSION

This work defines a model and describes a method to evaluate a site usability through its log comprised of the iterations of the user with the system.

The definition of a structured log containing both user's navigations with the system and the identification of the navigation strategy patterns makes evaluating the system usability easier. Thus, the evaluation becomes less dependent on the system. Moreover, the knowledge of the strategy patterns used allows us to formulate hypotheses on a user's intentions when he managed to obtain a solution to solve his task [7]. With a hypothesis on the intention it is possible to explain in order to guide to good solutions the tasks similar to those solved formerly; recover states that lead to bad solutions so that they can be avoided; customize the sequences of actions that are frequent facilitating their use with specific resources of interface.

The model described in this work automates the evaluation of a WEB site's navigation, reducing the time required to obtain the evaluation results as human intervention in this process is also reduced. The model also increases system usability, as the user will be able to complete their task quicker, not wasting time seeking the most adequate strategy toward their problem, and will reach better results, not taking paths that would lead to erroneous solutions.

REFERENCES

1. Bauer, C. and Scharl, A. Advanced design of Web information systems based on dominant and emerging web communication patterns, in *Proceedings of AusWeb'99, Fifth Australian World Wide Web Conference*, (Lismore Australia, 1999), Southern Cross University.
2. Castilho, J., Hartson, H. and Hix D. *Remote Usability Evaluation at a Glance*. 1997.
3. Barret, R., Maglio, P.P. and Kellem, D.C. How to *Personalize the Web*. IBM Almaden Research Center, San Jose CA, 2001.
4. Rossi, G., Schwabe, D. and Guimarães, R.M. Designing Personalized Web Applications, in *Proceedings of ACM'01* (Hong Kong, May 2001), ACM Press.
5. Kobsa, A. Generic User Modeling Systems. *User Modeling and User-Adapted Interaction 11* (2001), 49-63.
6. RiskControl. Available at www.riskcontrol.com.br/.
7. Pazzani, M. J. *Learning from historical precedent*. University of California, 1991.