Log Analysis of User Behaviour in the Renardus Web Service

Traugott Koch (1,2), traugott.koch@lub.lu.se; Anders Ardö (2), Koraljka Golub (2),
{anders.ardo|koraljka.golub}@it.lth.se
1 Knowledge Technologies Group, NetLab, Lund University Libraries, Sweden
2 Knowledge Discovery and Digital Library Research Group (KnowLib), Department of Information Technology, Lund University, Sweden

1. INTRODUCTION

Renardus (http://www.renardus.org) is a distributed Web-based service, which provides integrated searching and browsing access to quality controlled Web resources from major individual subject gateway services across Europe (funded by the EU's Information Society Technologies 5th Framework Programme until 2002). Navigation features are, among others, simple and advanced search, and subject browsing. Browsing is based on intellectual mapping of classification systems used by the distributed gateway services to the Dewey Decimal Classification (DDC). In addition to the dominating hierarchical directory-style of browsing (Gen. Browse), there are several other supporting features: graphical fisheye presentation of the classification hierarchy (Graph. Browse), search entry into the browsing structure (Search Browse) and merging of results from individual subject gateways (Merge Browse). Fig. 1 shows the main features, indicating their share of the activities (circle sizes) and transitions (arrow sizes) (only values above 1% are displayed):

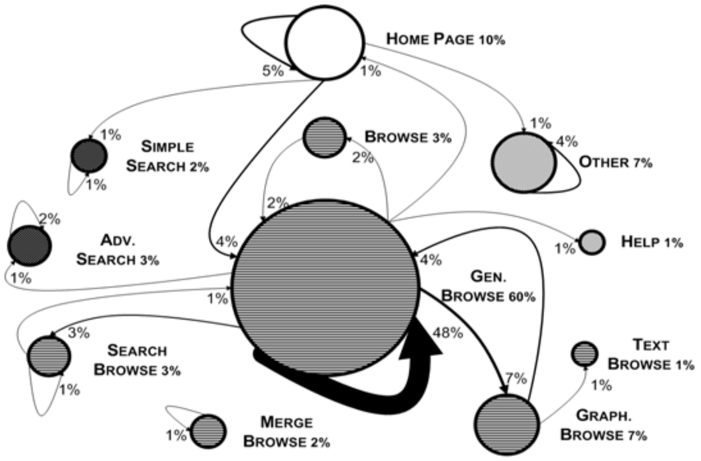

Fig. 1

With the overall purpose of improving the Renardus Web service, the research aims to study:

- the detailed usage patterns (quantitative/qualitative, paths through the system)
- the balance between browsing and searching or mixed activities
- typical sequences of usage steps and transition probabilities in a session
- typical entry points, referring sites, points of failure and exit points
- the usage degree of the browsing support features.

2. APPROACH

The Renardus project did a limited human evaluation of the service. Because of the high cost of full usability lab studies, we wanted to explore to what a degree a thorough log analysis – monitoring unsupervised usage – could provide valuable insights and working hypotheses as the basis for good usage and usability studies. Many sources of problems might be discovered already at this stage. A thorough log analysis requires several steps, starting with cleaning the log files with regard to activities from search engines, crackers, local administration, images etc. More than 2.3 million Renardus log entries boiled down to 630,000 user entries. The second step, based on heuristics, was to remove further 80,000 entries as probable machine activities. In order to study behaviour we needed to group log entries into user sessions. The basis for our further analysis turned out to be 155,000 user sessions, corresponding to 550,000 log entries, spanning over the period of 16 months. Each entry was classified into one of eleven different activities offered by Renardus. These activities were then used to characterize user behaviour, via a typology of usages and sequences.

3. PRELIMINARY FINDINGS

The log files analyzed show global usage of Renardus from about 99,605 unique machines and 351 unique top-level domains. First figures indicate that about 13% of our unique user machines have been returning to the service, which is a comparably good value for "faithful" users.
The levels of usage of the main Renardus features are highly uneven (cf. Fig. 1). The most surprising finding is the clear dominance of browsing activities (80%). This is a highly unusual ratio compared to other published evaluations and common beliefs. Among possible reasons are: a) the fact that 71% of the users reach browsing pages directly via search engines (Google and Yahoo! dominating); b) the layout of the home page focuses on browsing (22% of all users enter Renardus at the home page/the "front door" of the service).
Users tend to stay in the same feature (e.g. Adv. Search) and group of activities, whether it is browsing, searching or looking for background information, despite the provision of a full navigation bar on each page of the service. Especially the transitions between browsing and searching activities are less frequent than expected and hoped for. Fig. 2 demonstrates this by displaying the main transitions from each feature to other features of the service (the percentages – above 5% – displayed with the arrows relate to the feature they originate from).

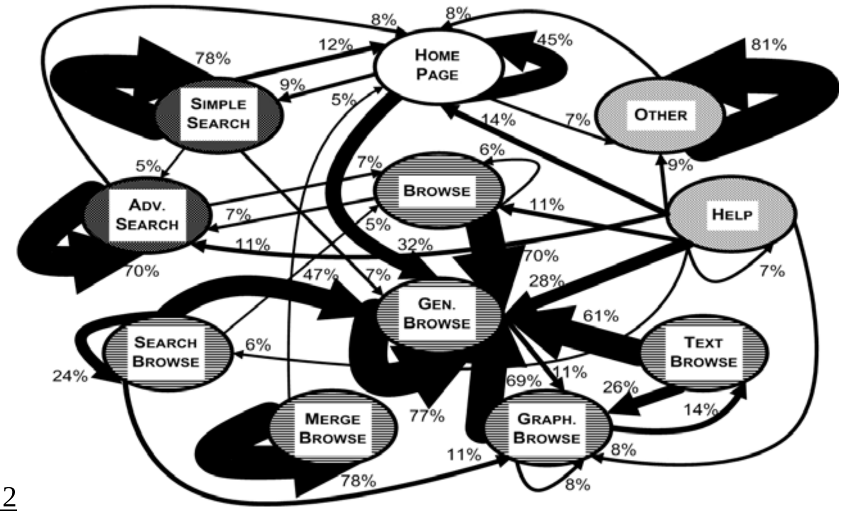

Fig. 2

Services like Renardus need to be designed for receiving the user where she first enters the system and provide search strategy support for the full usage of the system's features. The special browsing support features of the service are quite well used and worthwhile to further develop. Many users employ a surprisingly rich variety of navigation and browsing sequences and often alternate between many different features. For example, one session has the following sequence (the numbers indicate the repeated usage of the same feature):
home 2 - genbrowse 3 - browse 1 - home 2 - html 3 - genbrowse 6 - graphbrowse 1 - genbrowse 1 - graphbrowse 1 - genbrowse 1 – graphbrowse 1 - textbrowse 1 - graphbrowse 1 - genbrowse 4 - graphbrowse 1 - searchbrowse 2 - graphbrowse 1 - advsearch 1 - graphbrowse 1 - browse 1 - genbrowse 2 - graphbrowse 1 - textbrowse 1 - genbrowse 3 - advsearch 1 - showadvsearch 2 - scan 1 - showadvsearch 1 - scan 2 - advsearch 1 - showadvsearch 1 - browse 1 - genbrowse 1.

Directory-style of browsing in the DDC-based browsing structure is the clearly dominating activity in Renardus (60%). We found surprisingly long unbroken sequences of up to 90 steps in the DDC directory trees, even if the clear majority limits themselves up to 10 such steps (cf. the detailed view in Fig. 3b).

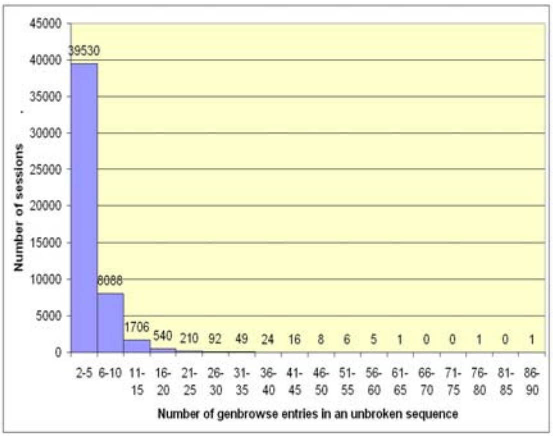

Fig. 3a

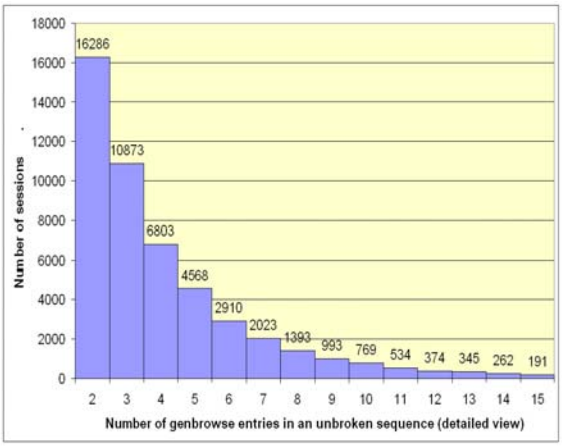

Fig. 3b

Use of the graphical DDC browsing overview is the second most frequent activity in Renardus, after the directory-style browsing. In 11% of the cases, directory-style browsing has been followed by the usage of the graphical overview.

Analysis of the popularity of DDC sections and classes and the navigation behaviour of users in the DDC structure allow good insights into distribution of topical interests and into the suitability of DDC system and vocabulary.

Systematic browsing of large information systems with the help of classification hierarchies seems to be widely accepted by users, especially when there is graphical support.

4. FUTURE WORK

These findings indicate that a thorough log analysis can provide deeper understanding of how the service really works and can be improved and they might offer useful hypotheses for advanced user studies.

Future work aims at investigating questions like:

- are there stable usage and browsing patterns and different behaviours of specific user groups?
- to what a degree is the actual design of the system influencing user behaviour, especially with regard to the different usage level of browsing versus searching activities?
- how can we provide search strategy support and improve the support for systematic browsing of large subject structures?

In order to make up for shortcomings of the log analysis approach, the following investigations will be needed:

- use cookies to identify the pages outside Renardus users explore as a result of Renardus navigation
- evaluate user behaviour in supervised sessions/usability lab
- evaluate the accuracy and success of Renardus to help answering user questions.

REFERENCES

1. Renardus Home Page.
   http://www.renardus.org
2. Presentation of some more detailed Renardus log analysis results. http://www.it.lth.se/knowlib/renardus-log/log-analysis.html
3. Koch, T., Neuroth, H., & Day, M. (2001). Renardus: Cross-browsing European subject gateways via a common classification system (DDC). In *Proceedings of the IFLA Satellite Meeting on Subject Retrieval in a Networked Environment*, 14-16 August 2001, Dublin, OH, USA. UBCIM Publications - New Series Vol. 25, München, 2003, 25-33. Manuscript at:
   http://www.lub.lu.se/~traugott/drafts/preifla-final.html
4. Hollman, J., Ardö, A. & Stenström, P. (2002). Empirical observations regarding predictability in user access behaviour in a distributed digital library system. In *Proceedings of the 16th International Parallel and Distributed Processing Symposium*, IEEE, April 2002, 221-228.

Acknowledgment

The Swedish Agency for Innovation Systems provided funding for this research.