



LUND UNIVERSITY

Electrical and Information Technology

LUP

Lund University Publications

Institutional Repository of Lund University

Found at: <http://www.lu.se>

This is an author produced version of the paper published in
Cataloging & Classification Quarterly

This paper has been peer-reviewed but does not include the
final publisher proof-corrections or journal pagination.

Citation for the published paper:

T. Koch, K. Golub, A. Ardö: *Users browsing behaviour in
a DDC-based Web service: A Log Analysis*, Cataloging &
Classification Quarterly, Vol. 42, No. 3/4, pp. 163-186, 2006.

Access to the published version may require subscription.
Published with permission from: The Haworth Press

Users Browsing Behaviour in a DDC-Based Web Service: A Log Analysis

Traugott Koch¹, Koraljka Golub¹, Anders Ardö^{1*}

Abstract

This study explores the navigation behaviour of all users of a large web service, Renardus, using web log analysis. Renardus provides integrated searching and browsing access to quality-controlled web resources from major individual subject gateway services. The main navigation feature is subject browsing through the Dewey Decimal Classification (DDC) based on mapping of classes of resources from the distributed gateways to the DDC structure.

Among the more surprising results are the hugely dominant share of browsing activities, the good use of browsing support features like the graphical fish-eye overviews, rather long and varied navigation sequences, as well as extensive hierarchical directory-style browsing through the large DDC system.

Keywords

Browsing behaviour; Log analysis; User studies; DDC; Dewey Decimal Classification; Subject gateways; User interfaces

¹ Knowledge Discovery and Digital Library Research Group (KnowLib), Lund University, Sweden. traugott.koch@lub.lu.se

*Acknowledgment

The Swedish Agency for Innovation Systems provided the main funding for this research. This work was partially funded by European Union (EU) under project ALVIS – Superpeer Semantic Search Engine (EU 6. FP, IST-1-002068-STP). This work was partially funded by DELOS – Network of Excellence on Digital Libraries (EU 6. FP IST, G038-507618).

INTRODUCTION

As many research communities are increasingly concerned with issues of interaction design, one of the current foci in information science is on user behaviour in seeking information on the World Wide Web. A frequently applied methodology for studying this behaviour is log analysis. This approach has several advantages: users do not need to be directly involved in the study, a picture of user behaviour is captured in non-invasive conditions, and every activity inside the system can be tracked.

User log studies mainly use the average analytical approaches of existing software packages for statistical reporting. Such software provides limited knowledge of user behaviour¹, since it only produces comparatively general insights into aspects of information services, such as number of users per month or the mostly followed hyperlink, and thus tells little about specific navigation behaviour.

A variety of aspects of user information-seeking behaviour using log analysis have been studied previously, in digital libraries², web search engines^{3,4,5}, and other web-based information services. Browsing behaviour has not been studied that much.

The common belief seems to be that users prefer searching to browsing: Lazonder⁶ claims "...students strongly prefer searching to browsing". Jacob Nielsen⁷ states the following: "Our usability studies show that more than half of all users are search-dominant, about a fifth of the users are link-dominant, and the rest exhibit mixed behaviour. The search-dominant users will usually go straight for the search button when they enter a website: they are not interested in looking around the site; they are task-focused and want to find specific information as fast as possible. In contrast, the link-dominant users prefer to follow the links around a site: even when they want to find specific information, they will initially try to get to it by following promising links from the home page. Only when they get hopelessly lost will

link-dominant users admit defeat and use a search command. Mixed-behaviour users switch between search and link-following, depending on what seems most promising to them at any given time but do not have an inherent preference”.

These observations have implications for building searching-oriented user interfaces. However, those results could be dependent on a number of issues that might have not yet been recognized. One such issue is, for example, the role of the web page layout in “favouring” either of the two strategies. Hong⁸ conducted a study on browsing strategies and implications for design of web search engines. The study reports that existing browsing features of search engines are insufficient to users. Even within the Renardus project, an initial belief about potential user requirements was that end-users preferred searching to browsing⁹. After the browsing interface had been built, it showed that browsing was much favoured.

The overall purpose of our project was to gain insights into real users navigation and especially browsing behaviour in a large service on the web. This knowledge could be used to improve such services, in our case the Renardus service¹⁰ which offers a large DDC browsing structure. Renardus 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. (The Renardus project was funded by the European Union’s Information Society Technologies 5th Framework Programme until 2002).

The research aimed at studying the following topics: the unsupervised usage behaviour of all Renardus users, complementing the initial Renardus user enquiry; detailed usage patterns (quantitative/qualitative, paths through the system); the balance between browsing, searching and mixed activities; typical sequences of user activities and transition probabilities in a session, especially in traversing the hierarchical DDC browsing structure; the degree of usage of the browsing support features; and typical entry points, referring sites,

points of failure and exit points. Because of the high cost of full usability lab studies, we also wanted to explore whether a thorough log analysis could provide valuable insights and working hypotheses as the basis for good usage and usability studies at a reasonable cost.

The paper provides short background information about Renardus (I. Background); the methodology applied in this study is described in section two (II. Methodology); the analysis, hypotheses and results regarding the general usage of Renardus, the browsing behaviour and the usage of the DDC are presented in the third section (III. Results). A summary of the results and some ideas for further investigation conclude the paper (IV. Conclusion).

I. Background

Renardus service

Renardus¹⁰ exploits the success of subject gateways, where subject experts select quality resources for their users, usually within the academic and research communities. This approach has been shown to provide a high quality and valued service, but encounters problems with the ever increasing number of resources available on the Internet. Renardus is based on a distributed model where major subject gateway services across Europe can be searched and browsed together through a single interface provided by the Renardus broker. The Renardus partner gateways cover over 80,000 predominantly digital, web-based resources from within most areas of academic interest, mainly written in English.

The Renardus service allows searching several subject gateways simultaneously. What is searched are “catalogue records” (metadata) of quality controlled web resources, not the actual resources. There are two ways to search the service, either through a simple search box that is available on the Renardus “Home” page or through the “Advanced search” page

allowing combination of terms and search fields and providing options to limit searches in a number of different ways. A pop-up window of a list of words alphabetically close to the entered word (for title, DDC, subject and document type) supports the search term selection.

Apart from searching, Renardus offers subject browsing in a hierarchical directory-style (cf. Technology | Agriculture | Animal husbandry¹⁴). It is based on intellectual mapping of classification systems used by the distributed gateway services to the DDC. There are also several browsing-support features. The graphical fish-eye display presents the classification hierarchy as an overview of all available categories that surround the category one started from, normally one level above and two levels below in the hierarchy. This allows users to speed up the browsing and get an immediate overview of the relevant Renardus browsing pages for a subject. The feature “Search entry into the browsing pages” offers a short-cut to categories in the browsing tree where the search term occurs. The lower half of the browsing pages, as a result of the classification mapping, offers the links to the “Related Collections” of the chosen subject. In case users do not want to jump to the parts of the gateways offering related collections, an option of Merging the resource-descriptions from all related collections is available.

For a more detailed description of Renardus, see, for example Koch, Neuroth, and Day¹¹. All related publications are given at the web page “Project Archive and Associated Research and Development”¹².

II. Methodology

Before Renardus was finally released and the EU project concluded in 2002, an end user evaluation of the Renardus pilot subject gateway¹³ was carried out during Fall 2001 which led to some service improvements. The results and shortcomings of this initial user

study stimulated us to try the full study of Renardus user logs which is presented in this paper.

Log analysis was chosen because it costs considerably less than full usability lab studies and has the advantage that it is an unobtrusive means of capturing unsupervised usage. This thorough log analysis required several steps which are described below: cleaning of the log files, defining of user sessions, categorization into activity types and the creation of datasets and structures to allow the creation of statistics and the testing of hypotheses.

Cleaning the log files

The log files used spanned 16 months between summer 2002 and late fall 2003. They first had to be cleaned from entries created by search engine robots, crackers (users performing unauthorized activities), local administration, images, etc. The largest group of removed entries, almost half of all log entries, was that containing images and style sheets (1,107,378). Further, 516,269 entries were removed because they originated from more than 650 identified robots, and an additional 12,647 entries because they were from crackers. Various other entries not relating to real usage of Renardus for information seeking, e.g. 17,586 redirections, about 9,000 local administrative activities, error codes and HTTP head entries, had to be removed.

Thus, in the first step, the total number of 2,299,642 log entries was reduced to 631,711 entries. From this dataset only some general Renardus usage statistics was derived. For the analysis of real user behaviour in Renardus, several further steps and separate datasets were required.

Defining sessions

After cleaning, the log all entries were grouped into user sessions. A session was heuristically defined as containing all entries coming from the same IP address and a time gap of less than one hour from the prior entry from the same IP-address.

Defining activity types

Each log entry was classified into one of eleven different main activities offered by Renardus. These activities were then used to characterize user behaviour, via a typology of usages and sequences of activities.

Browsing activities:

“Gen. Browse”, hierarchical directory-style browsing of the DDC (cf. e.g. ¹⁴); “Graph. Browse”, graphical fisheye presentation of the classification hierarchy (cf. e.g. ¹⁵); “Text Browse”, text version of the graphical fisheye presentation; “Search Browse”, search entry into the browsing structure; “Merge Browse”, merging of results from individual subject gateways; “Browse”, DDC top level browsing page on the home page.

Searching activities:

“Simple Search” with “showsimpsearch” for result display; “Adv. Search”, advanced search with “showadvsearch” for result display and “scan” for scanning certain data indices.

Other activities:

“Home Page”; “Help”; “Other” other informational pages, including project documentation.

Creating datasets for studying information-seeking behaviour

To try to make sure that we studied only human behaviour in Renardus, we removed, in a further step, another 82,490 entries judged as probable machine activities. This determination was based on heuristic criteria, for example, all sessions containing only one entry; sessions shorter than two seconds.

Most of the analysis in this paper regarding human activities in Renardus is based on a dataset containing 464,757 entries grouped into 73,434 user sessions. Only in a few calculations (especially in the section “Browsing sessions”) we used a further subset of this dataset. The different datasets were stored in a relational database and SQL was used to query them to create statistical tables and to test various hypotheses against the log file data.

III. Results

Global usage

Renardus was accessed from 99,605 unique machines (IP-numbers) during the 16 month period studied. With 351 unique top-level domains or countries identified (a considerable part of the IP-numbers could not be identified), it is apparent that Renardus has a truly global audience. IP-numbers from the USA topped the list with about 30%, other .net and .com domains followed with 8-10%. Renardus Project partner countries were led by Finland with 5%. Canada, Australia, the Philippines, Italy and India were other countries exceeding 1% of the IP-numbers.

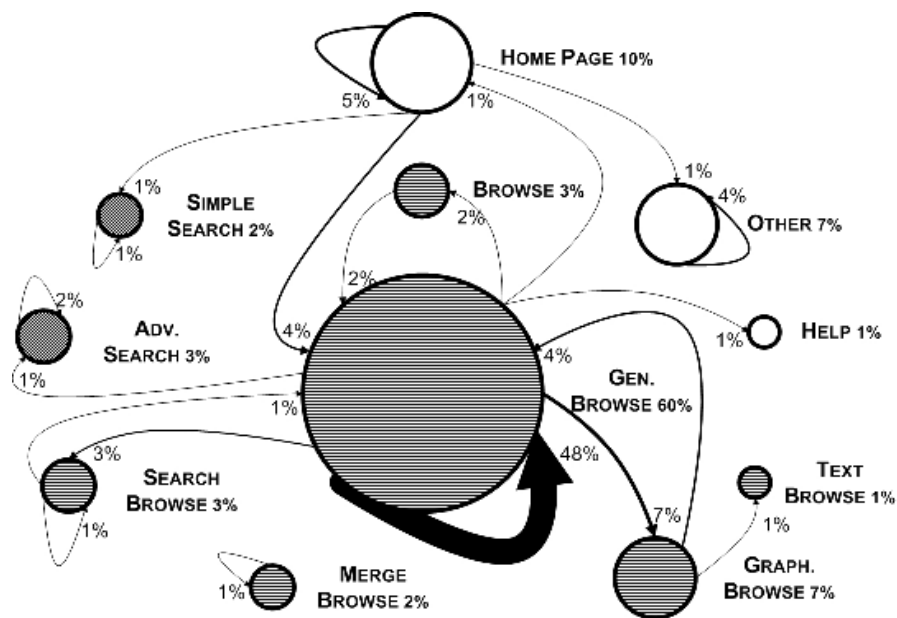
The user sessions are of considerable length: 33% are longer than 2 minutes and 10% are longer than 10 minutes. The time users might have been exploring participating gateways after leaving Renardus is not included.

The figures indicate that more than 851 different hosts referred users to Renardus. As much as 56% of all referred sessions came from various Google servers and 24% from Yahoo!

Renardus seemed to be able to attract and keep many “faithful” users during the first 16 months after release. 13% of all unique user machines were returning to the service, which is a comparatively good value.

A Information seeking activities

Main activities, transitions



[Figure 1 Main Renardus features, indicating their share in all activities, and major transitions between the activities.]

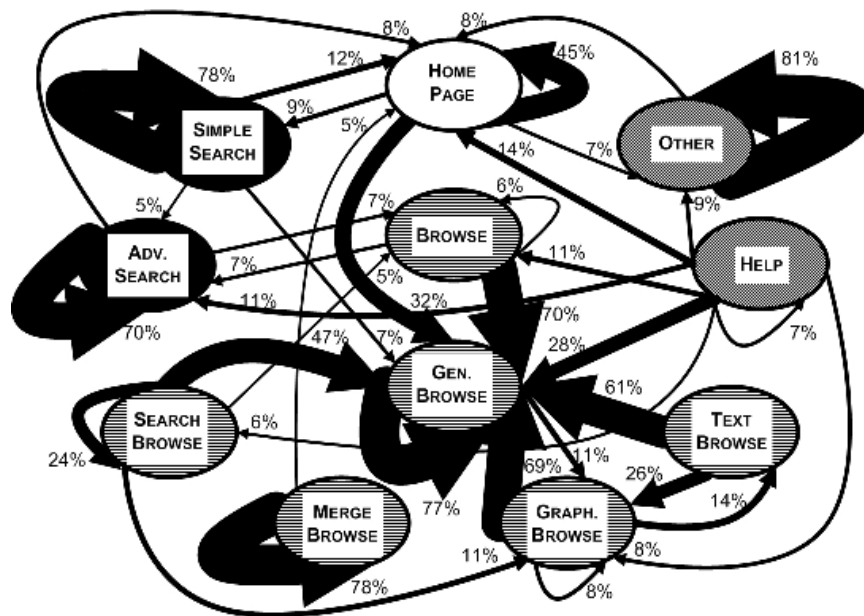
Figure 1 illustrates the share of each activity and transition in the following ways: the share of each of the main activities is indicated by the circle size; and the share of the major transitions between different activities is indicated by arrow size. Only values above 1% are displayed. It shows that 60% of all Renardus activities are directory-style browsing using the

DDC structure (Gen. Browse; for the abbreviations used here and throughout the paper, cf. the description under II Methodology: Defining activity types). 48% of all transitions in Renardus are steps from one such topical page/DDC class to another.

The four special browsing support features are comparatively well used. As many as 45% of the sessions dominated by browsing use two or more different types of browsing activities. As many as 14% use three to five different types (see Browsing sessions below).

Use of the graphical DDC browsing overview (Graph. Browse) is the second most frequent activity in Renardus (7%), after the directory-style browsing. The transition from the dominant directory browsing in the DDC structure to a graphical display is clearly the largest single transition in Renardus, after subsequent directory browsing steps.

Related to Gen. Browse, in 11% of the cases, directory-style browsing has been followed by the usage of the graphical overview (see Figure 2). For further reasoning about these findings see below.



[Figure 2 Transition probabilities (more than 5%-transitions only).]

Figure 2 illustrates another important finding. Users tend to stay in the same feature and group of activities, whether it is a single activity like Gen. Browse or a group like browsing, searching or looking for background information, despite the provision of a full navigation bar on each page of the Renardus service. In particular, the transitions between browsing and searching activities are less frequent than expected and hoped for. Figure 2 demonstrates this by displaying the main transitions from each feature to other features of the service (the percentages displayed close to the arrows relate to the feature they originate from). For example, 77% of all transitions from one Gen. Browse activity are directed to another Gen. Browse activity and 11% to Graph. Browse.

As the early user study in 2001 showed (in table 18 in the study)¹³, the Renardus pilot service was mostly considered very easy or easy to navigate already, although a fifth of the respondents found navigating through the different parts of the service difficult or very difficult. We conclude that advanced online services need to provide some kind of search strategy support. They need to be designed for receiving the user where he/she first enters the system and to assist with user navigation through the whole system with more than a ubiquitous navigation bar (which is offered by Renardus on all pages).

General navigation sequences

Many users engage in several different activities during their session: about 46% in one activity, 20% in two, 16% in three different activities. About 18% of the user sessions have between 4 and 11 different activities.

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 first number indicates the number of immediate repetitions of the same feature, the second gives the length of this activity in seconds).

home 3 3; genbrowse 4 31; graphbrowse 1 1; genbrowse 3 17; home 1 1; browse 1 1; genbrowse 2 3; searchbrowse 1 1; genbrowse 7 152; searchbrowse 1 1; genbrowse 4 24; help 1 1; home 1 1; genbrowse 4 29; graphbrowse 1 1; searchbrowse 1 1; genbrowse 1 1; searchbrowse 1 1; genbrowse 1 1; browse 1 1; genbrowse 3 2; graphbrowse 1 1; genbrowse 2 2; home 2 2

When we look at the most frequent sequences of activity types (immediate repetition of the same type is not counted), we find 4,810 different sequences. The top ten sequences are presented in Table 1. The most frequent sequences, apart from mergebrowse and showsimpsearch, are (in and) between browsing activities.

Type of activity	Sessions	%
(repetitions of) genbrowse	30,606	41.7%
other	7,403	10.1%
genbrowse-graphbrowse-genbrowse	3,860	5.3%
genbrowse-graphbrowse	3,590	4.9%
genbrowse-searchbrowse	2,812	3.3%
(repetitions of) mergebrowse	2,391	3.3%
(repetitions of) showsimpsearch	1,705	2.3%
genbrowse-browse-genbrowse	1,635	2.2%
genbrowse-searchbrowse-genbrowse	1,236	1.7%
genbrowse-browse	1,035	1.4%
all less frequent sequences		23.8%

[Table 1 Most frequent sequences of activity types.]

When we look at a more detailed table of sequences including immediate repetitions of the same activity (not reproduced here), the dominance of browsing and the very high

number of variations in navigation is well illustrated. In 73,434 user sessions we find as many as 16,377 different sequences; however, the top 10 most frequent sequences (with more than 1,000 instances each) cover 41.7% of all sessions. In the top 6, and numbers 9-11 among the 11 most frequent sequences, the user exclusively repeats the same activity. Only numbers 7 and 8 involve a switch between different activities (from genbrowse to graphbrowse and from genbrowse to searchbrowse). In the five most frequent cases genbrowse is the repeated activity. The sequences where only the same activity type is repeated cover about 50% of all sessions. This further underlines our earlier finding that a surprisingly large part of the users stay in the same (group of) activities.

Browsing vs. searching

The levels of usage of the main Renardus features are highly uneven (cf. Figure 1). The most surprising finding is the clear dominance of browsing activities, about 80%. Depending how “dominance of browsing” is defined: 76% of all activities are browsing; 80.5% of all sessions are dominated by browsing. Searching has a much lower share, between 3 and 6%.

This is a highly unusual ratio compared to other published evaluations and common beliefs (cf. Introduction). A possible reason is that most of the browsing pages are indexed by search engines. 71% of the users reached browsing pages directly via search engines and start their Renardus navigation at a browsing page. Taken together with the clear tendency to stay in the same (group of) features, these facts “favour” browsing. Additionally, the layout of the home page invites browsing by putting the browsing structure on top of the search box. Still, among users starting at the home page, 57% browse and only 12.5% search (only 22% of all users enter Renardus at the home page/the “front door” of the service, however).

In spite of the dominance of browsing and the tendency to stay in the same group of activities, we see a certain amount of switching between browsing and searching during the same session. In as few as 7.3 % of all sessions users switch between a browse and a search activity, out of which 4.5% of sessions have one switch, 1.9% have two, 0.4% have three, and 0.5% have more than three switches.

The largest number of switches per session is 20. Out of 27 different kinds of switches between browsing and searching, 7 start with a search. Switching from browsing to searching is much more frequent than the opposite. Users at the search pages need to be pointed to the benefits of browsing.

Browsing sessions

For the calculations in this section we use a subset of our usual dataset, containing 378,267 entries in 58,954 user sessions, defined by a share of more than 50% browsing activities: sessions where “browsing is dominant”.

The shares of sessions with a certain number of different activities are almost the same as for all Renardus sessions (cf. the beginning of General navigation sequences). So, even sessions with dominant browsing show as much variety in activities as most other sessions.

Many browsing sessions use more than one type of browsing activity, including the browsing support features: Graph. Browse, Text Browse and Merge Browse. As many as 45% of the sessions dominated by browsing show two or more browsing activities and 14% three to five different types of browsing. We find up to 95 individual browse activities per session, with gracefully degrading numbers from two activities and down.

Two different groups of users

Because of the big influence of referrers like search engines, 71% of the human user sessions start at browsing pages pointed to by referrers, whereas 22% start at the homepage (16,300 out of 73,434 sessions). This quantitatively surprising result stimulated us to check if these two “groups” of users show significantly different navigation behaviour. Sessions starting at home have almost twice as many entries per session than sessions starting elsewhere (10 vs. 5.8 entries per session; 35.8% of all entries). Thus, home starters carry out many more activities per session than the other user group.

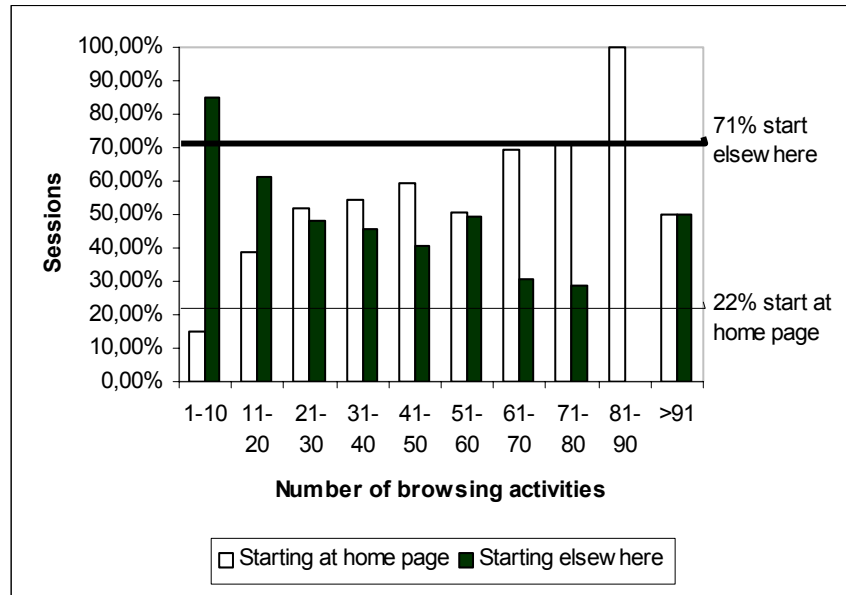
Type of activity	Starting at home		Starting elsewhere	
	Entries	%	Entries	%
Browsing	94,215	56.6	259,471	87.0
Searching	20,831	12.5	8,099	2.7
Other	51,139	30.9	30,684	10.3
Total	166,503		298,254	

[Table 2 Types of activities for the two different groups of users.]

Users jumping into the middle of the Renardus service are carrying out browsing activities in 87% of all cases and only 2.7% searching activities (Table 2).

Users starting at the Renardus homepage/“frontdoor” show a level of browsing of almost 57%, and 12.5 % searching. Three times as often they visit other pages and five times as often search pages compared to the other group. These are probably the users who go deliberately to Renardus, whereas a large part of users starting elsewhere, most often in the browsing pages, end up there “ignorantly” after a search in a search engine. The latter overwhelmingly stay in the browsing activities.

People starting elsewhere have a much higher percentage of browsing among their activities. Home starters, however, do considerably more browsing activities compared to their share of all sessions: 53.2% of the sessions show more than 11 browsing activities and 36.8% more than 30 browsing activities.



[Figure 3 Browsing activities of the two groups of users]

Figure 3 shows that the home starters clearly dominate the sessions with many browsing activities. A more detailed analysis shows that they are active in browsing activities to a higher and increasing degree starting with 8 browsing activities, compared with their share in all sessions (21%). Quite the opposite is true for users starting elsewhere. They are overrepresented up to the level of nine browsing activities with an ever-decreasing tendency.

Home starters also exceed their share when it comes to the number of different activity types, all types are counted (in browsing sessions) except when there are three different activities. From five different activities and higher, they have more than twice their share and dominate clearly.

When it comes to the number of different browsing types (in browsing sessions), home starters exceed their share when it comes to carrying out between three and five different browsing types.

B DDC USAGE

DDC analysis

Analysis of the popularity of DDC sections and classes and the navigation behaviour of users in the DDC structure allow good insights into the distribution of topical interests and the suitability of the DDC system and vocabulary. The findings from the log analysis can, however, only help create hypotheses and need to be complemented by investigative sessions with the users. The most frequently used parts of the DDC hierarchy at the top hierarchical level are given in Table 3.

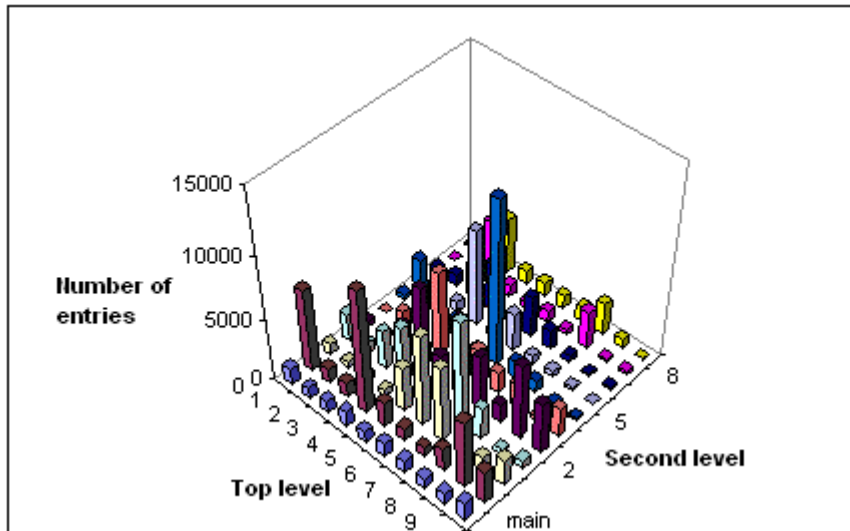
Entries	DDC	Class
50,784	3	Social sciences
46,209	5	Science
30,955	6	Technology
26,015	2	Religion
22,081	7	Arts & recreation
17,994	8	Literature
16,828	9	History & geography
16,527	0	Computers, information & general reference
13,839	4	Language
13,428	1	Philosophy & psychology

[Table 3 Most frequently used parts of the DDC hierarchy at the top hierarchical level.]

All DDC classes show generally good usage levels (users jumping to one class and not continuing browsing are not counted). Compared to what one would expect in a global internet setting, Religion ranks surprisingly high and Computers etc. unexpectedly low (see

Table 3). Here the vocabulary used in the DDC captions could play a role, e.g. many computing-related terms used in Internet searching do not directly occur in the captions.

On the second hierarchical level, surprisingly large topical areas are Christian denominations (DDC 28), German & related literatures (83), Social problems (36) and Earth Sciences (55; cf. Figure 4).



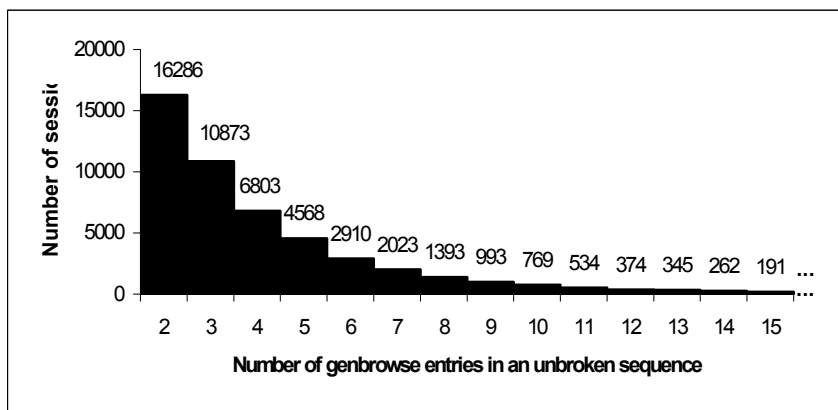
[Figure 4 Most frequently used parts of the DDC hierarchy at the second hierarchical level.]

Unexpectedly frequent visits to individual topics like 552.1 Igneous rocks (the sixth most visited individual page with 2,436 directory browsing activities) could be due to the fact that little information might be found about such a concept in the search engines or to the fact that other sites made prominent links to this topic page in Renardus.

Directory style of browsing in the DDC hierarchy

The directory-style of browsing in the DDC-based browsing structure is clearly the dominant activity in Renardus (about 60%). 67% of all browsing activities are DDC directory

browsing (254,660 out of 378,264 entries in browsing sessions). Two thirds of the latter (167,628) appear in unbroken sequences. In these cases, not even browse support features are used between directory browsing steps. While the clear majority of users limit themselves to 10 or fewer steps (for distribution see the Figure 5), we found surprisingly long unbroken browsing sequences of up to 86 steps in the DDC directory trees.



[Figure 5 Number of genbrowse activities in sessions (up to 15).]

These are very unexpected results. People looking for information on the web are often said to use as few clicks as necessary, switching frequently to other services and activities, having a very short attention spans. Browsing the DDC hierarchies in a directory style of steps at such quantity and lengths is one of the most significant results of this log study.

Jumping in the DDC hierarchy

Since the DDC browsing area in the Renardus user interface displays the higher levels in the hierarchy in addition to the “parent” and the “child” classes, we can find out to what

degree users jump levels in the DDC hierarchy during unbroken directory browsing sequences.

Two of the support features, the graphical overviews and the “search entry to browsing pages”, were designed to relieve users from the “pain” of having to jump around in the hierarchy. Jumping one step up and another step down in the directory-style display is probably faster and easier than using the support features; moving farther away would possibly have been easier using the support features.

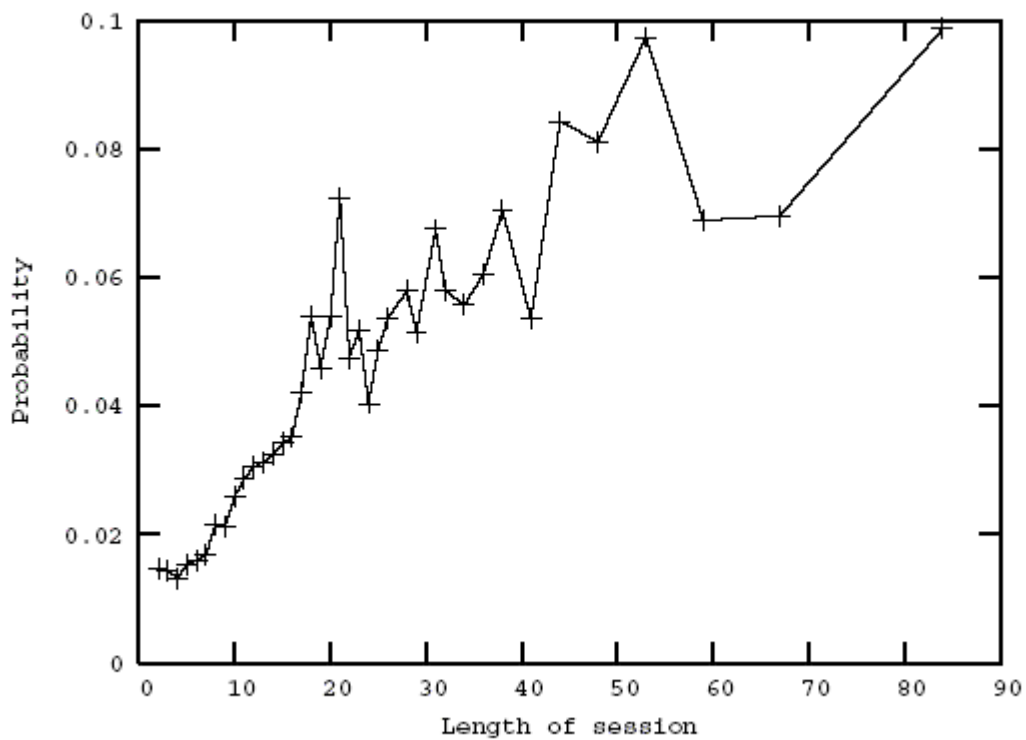
The following sequence is an example of a session featuring jumps within unbroken directory browsing:

start 62-; go to 624; go to 624.1; jump to 62-; go to 625; go to 625.1; go to 625;
go to 62-; go to 627; jump to 628; go to 628.1

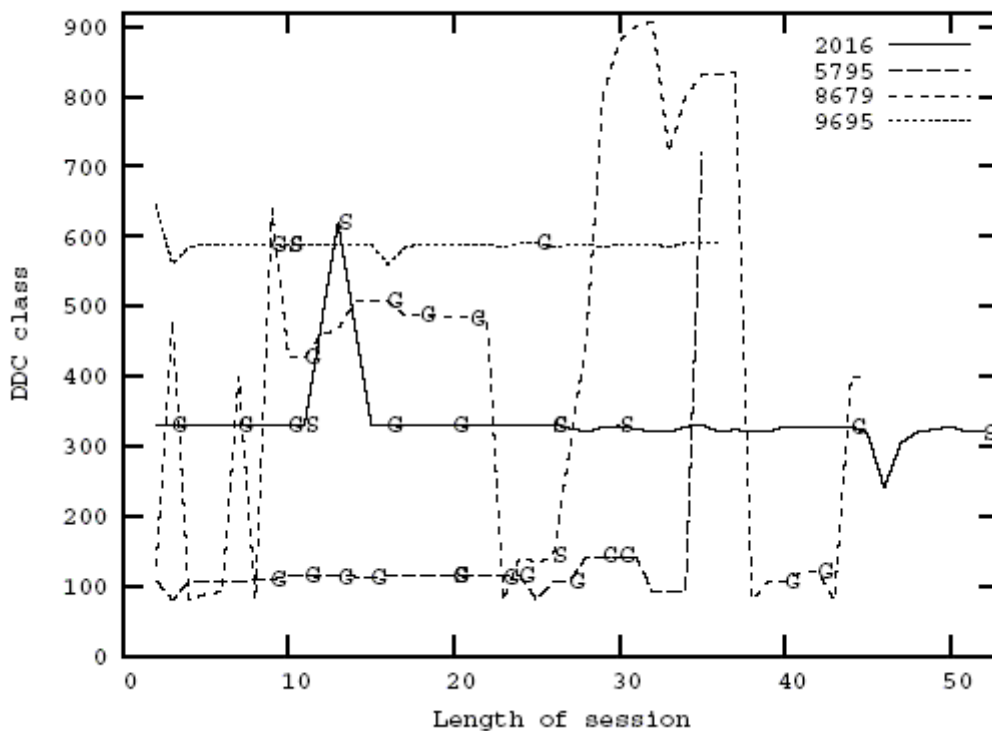
20.2% of all steps in sessions featuring unbroken directory browsing are jumps. Jumps occur in 40.8% of these sessions. In the sessions with jumps, on average 1.7 jumps are carried out. This is a decent number of cases but not excessively high. Many users make use of the support features, especially the graphical overviews, instead of jumping in the directory. This finding indicates at least, that the necessity to jump in the hierarchy is not putting off users.

As seen from Figure 6 the probability for a user in one session to browse in several main DDC classes increases with the length of the session. This might seem natural but it also implies that the longer the session, the shorter time spent within one main DDC class before moving to another. Each point in the figure is based on several sessions that together contain more than 2,000 browsing entries. Due to the heavy dominance of shorter sessions, the overall mean probability of moving between DDC main classes in a session is 3%.

Figure 7 shows a few individual sessions plotted with the number of browsing steps versus the visited DDC classes. For example, all classes within the “1—” branch of DDC are displayed between 100 and 200 on the vertical axis in such a way that the hierarchy is preserved, e.g. the closer two classes are in the hierarchy the closer they are plotted in the figure. Thus a horizontal line indicates that the user stays within a narrow area of DDC while vertical parts indicate jumps between different areas of DDC. The letter “G” indicates that the graphical overview was used while an “S” indicates that the search entry to the browsing structure was used at the indicated points in the sequence.



[Figure 6 Probability of moving between DDC main classes.]



[Figure 7 DDC browsing behaviour per session.]

Keywords and browsing

We wanted to find out whether the user managed to come close to his/her topic of interest when browsing DDC pages in Renardus. In order to get an indication of that, we compared the keywords entered by a given user into the search engine (Google) respectively entered into Renardus Search with the browsing pages visited subsequently by the same user.

The following list of examples shows keywords entered into Google and the => Renardus DDC class the user selected from the search result:

ancient continents => History of ancient world; of specific continents, countries, localities; of extraterrestrial worlds

perspective drawing => Drawing & decorative arts

“statistics of south america” => General statistics of specific continents, countries, localities in modern world

writing systems and etymology => Standard language--description and analysis

kinds of sedimentary rocks => Specific kinds of rocks

The sample studied showed very good hits in the Renardus DDC pages. Most queries matched terms in the DDC caption (which is also used as the title of the page), about 13% of the cases had partial hits there and partial matches in other class and directory “titles” mentioned on the same page (parent, child DDC classes; names of mapped directories from cooperating subject gateways). Most successful questions used 2-3 query terms; only 3% used one term.

It seems that a good initial hit is required to invite users to continue browsing in Renardus (the data here is derived from sessions containing more than one browsing activity).

The result says more about the search engine’s ranking algorithm (Google in the case of our sample) than about the Renardus pages and the suitability of the DDC captions. Part of the reason for the share of the good hits we see is the fact that only DDC pages from Renardus where the user’s search terms hits several words in the title or top half of the DDC page in Renardus have a chance to appear in the top of Google search results. Only such pages have a chance to be selected and visited by the user.

When checking queries and hits in Other Renardus pages (background and project information), we found great results, too: most hits seemed relevant and we couldn’t find many wrong hits on topical questions.

The following examples show search terms entered into Renardus Search (Q:) and DDC classes/pages used during browsing. Each case was gathered from the entire session and

was limited to sessions starting with Search and continuing with Browse activities. In the examples, queries and DDC captions are separated by a semicolon.

Q=chopin; vieuxtemps;

DDC=Arts & recreation; Music; Composers and traditions of music;

Q=paperin+valmistus; paperin+valmistus; papermaking; paper+technology;

DDC=Technology; Engineering; Engineering of railroads and roads; Engineering of railroads and roads; Engineering of railroads and roads; Railroads; Railroads; Astronautical engineering; Technology; Engineering; Engineering and allied operations; Engineering mechanics and materials; Science; Chemistry; Chemistry; Organic chemistry; Technology; Chemical engineering; Chemical engineering and related technologies; Biotechnology; Biotechnology; Pulp and paper technology; Genetic engineering; Electrical engineering; lighting; superconductivity; magnetic engineering; applied optics; paraphotic technology; electronics; communications engineering; computers; Electrical engineering; lighting; superconductivity; magnetic engineering; applied optics; paraphotic technology; electronics; communications engineering; computers; Electronics; Special topics; Optoelectronics; Pulp and paper technology; Conversion of pulp into paper, and specific types of paper and paper products; General topics; Properties, tests, quality controls;

The results of the evaluations of our sample remind us that users frequently follow more than one topic of interest during one session in an information system. In our sample 70% of all users seemed to pursue one topic in a session, 23% two topics, 2% three topics and 5% seemed to browse around without specific question. In some cases, topics looked for in

Renardus Search are not pursued when browsing, in other cases, a new topic (most often one) is investigated after the switch to browsing.

IV. Conclusion

The main purpose of this study was to explore the navigation behaviour of all users of a large web service, Renardus, using web log analysis, in order to improve the user interface and, especially, the browsing features of the system. In addition, we aimed at gaining some more general insights into users browsing and navigation in large subject classification structures, the benefits from system support and the problems and failures that occurred.

Our study indicates that a thorough log analysis can indeed provide a deeper understanding of user behaviour and service performance. Being an unobtrusive means of capturing unsupervised usage and offering a complete and detailed picture of user activities, log analysis can reveal quantitatively comprehensive, sometimes unexpected results, far beyond plain statistics.

In contrast to common beliefs, our study clearly indicates that browsing as an information-seeking activity is highly used, given proper conditions. About 80% of all activities in Renardus are browsing activities. A contributing reason to that dominance is the fact that a very high percentage (71%) of the users are referred from search engines or other linking sites directly to a browsing page in Renardus. The layout of the home page “invites” browsing, which certainly contributes to the fact that even users starting at the home page predominantly use the browsing part of the service.

Our study leads to a hypothesis which deserves further research: browsing is perceived as useful and dominates navigation in services similar to Renardus and under proper conditions.

The good use of the browsing support features, especially graphical overview and search entry to browsing pages, suggests that it would be worthwhile to further develop such support.

Since most visitors jump into the middle of the service, there might be a need to redesign the browsing pages so they would better serve as full-fledged starting points for comprehensive Renardus exploration. The ubiquitous navigation bar seems not sufficiently inviting. In making such changes, it would also be important to better understand the details of site indexing and ranking algorithms in search engines.

The study of navigation sequences shows that users employ a rich variety of navigation and browsing sequences, including rather long and highly elaborate paths through the system. Nevertheless, quantitatively dominating is, to a quite surprising degree, the tendency to stay in the same group of activities or individual activity, whether browsing, searching or background information. This finding points us to the importance of providing “search strategy” support to the users at the page where their actions take place.

From the behaviour as documented in the log files we could identify two clearly different groups of users: people starting at the homepage/frontdoor of the service (22%), and the majority of the users starting elsewhere. There are dramatic differences in their activity in the service. People starting at the homepage show almost twice as many activities per session, and use the non-browsing features three to five times as often. Their share of the browsing activities is smaller, but they primarily engage in the long sequences of browsing activities (8 and longer) and employ more types of browsing activities and more types of other activities in a session. The home page starters are seemingly a minority but represent high quality of usage of the service in a way the system designers have imagined and intended.

The DDC directory browsing is the single clearly dominating activity in Renardus (60%). Two thirds of it is done in unbroken directory browsing sequences. We see a

surprising average and total length of such browsing sequences, opposing the common belief of the short attention span of users of online services.

Thus, we get the surprising hypothesis that sequential, directory style of hierarchical (classification) browsing is found to be popular and useful in large services like Renardus, especially when there is graphical support.

Comparisons between search terms used and topics browsed indicated a very good chance to get relevant results from Renardus browsing when more than one search term was used. People using Renardus Search were capable of finding browsing pages corresponding to their queries. The system invited users to pursue more than one topic during a session.

Future work

Our findings indicate that log analysis has a clear potential as a method for studying information behaviour and the proper design of information systems. A lot could be gained from future work to investigate questions such as:

- To what degree does the actual design of the system influence user behaviour, especially with regard to the difference in usage levels of browsing versus searching activities?
- Can we identify additional specific usage and browsing patterns and different behaviours of specific user groups?
- What is the influence of the use of end-user adapted and multilingual DDC captions on browsing behaviour?
- How can we provide search strategy support and further improve the support for systematic browsing of large subject structures?

- What is the importance of the details of site indexing in search engines for the discovery of and navigation in large browsing systems?
- How can pages be redesigned so that they better serve as full-fledged starting points?

For more important results and improvements one would need to go beyond the log analysis and:

- evaluate user behaviour in supervised sessions/usability lab
- evaluate the accuracy and success of Renardus to help answering user questions
- use local URLs to identify what pages outside Renardus users explore as a result of Renardus navigation (links to participating subject gateways).

References

1. Harry Hochheiser and Ben Shneiderman. "Using interactive visualizations of WWW log data to characterize access patterns and inform site design." *Journal of the American Society for Information Science and Technology* 52, no. 4 (2001): 331–343.
2. S. Jones, S. J. Cunningham, R. McNab, and S. Boddie. "A transaction log analysis of a digital library." *International Journal on Digital Libraries* 3, no. 2 (2000): 152-169.
3. C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. "Analysis of a very large web search engine query log." In: *SIGIR Forum*, 33, no. 1(1999): 6-12.
<http://doi.acm.org/10.1145/331403.331405>

4. Seda Ozmutlu, Amanda Spink, and Huseyin C. Ozmutlu. 2004. "A day in the life of web searching: an exploratory study." *Journal of Information Processing and Management* 40, no. 2 (2004): 319-345. [http://dx.doi.org/10.1016/S0306-4573\(03\)00044-X](http://dx.doi.org/10.1016/S0306-4573(03)00044-X)
5. S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. "Hourly analysis of a very large topically categorized web query log." In: *Proceedings of the 27th annual international conference on Research and development in information retrieval, Sheffield, United Kingdom, 2004*, 321-328. <http://doi.acm.org/10.1145/1008992.1009048>
6. Ard W. Lazonder. "Principles for Designing Web Searching Instruction." *Education and Information Technologies* 8 (June 2003): 179–193. P. 181.
7. Jacob Nielsen. "Search and You May Find." Jakob Nielsen's Alertbox for July 15, 1997. <http://www.useit.com/alertbox/9707b.html>.
8. Xie Hong. "Web browsing: current and desired capabilities." In: *20th Annual National Online Meeting, 18-20 May 1999, New York, NY, US*, 523-37.
9. User requirements for the broker system: Renardus Project Deliverable D1.2. 2000. P. 23. http://www.renardus.org/about_us/deliverables/d1_2/D1_2_final.pdf.
10. Renardus Home Page. <http://www.renardus.org/>. [Since the project finished in 2002, this is only a demonstrator.]

11. Traugott Koch, Heike Neuroth and Michael Day. "Renardus: Cross-browsing European subject gateways via a common classification system (DDC). In: "Subject Retrieval in a Networked Environment", *Proceedings of the IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing and the IFLA Section on Information Technology, 14-16 August 2001, Dublin, OH, USA*, 25-33. München: UBCIM Publications New Series Vol. 25, 2003. Manuscript at: <http://www.lub.lu.se/~traugott/drafts/preifla-final.html>

12. Renardus Project Archive and Associated Research and Development. 2002.
http://www.renardus.org/about_us/project_archive.html.

13. User evaluation report: Renardus Project Deliverable D5.2. 2002.
http://www.renardus.org/about_us/deliverables/d5_2/D5_2_final.pdf.

14. Technology:Agriculture : page.
<http://www.renardus.org/cgi-bin/genDDCbrowseSQL.pl?ID=10191&node=AAZNG>

15. Graphical browsing page for Technology ...: Mining for specific materials.
<http://www.renardus.org/cgi-bin/imageDDCbrowseSQL.pl?node=ABDPH&ID=10193&pmat=N&pnavnode=Y&pgraph=matcirc>

All electronic resources have been accessed 20 January 2005.