Search Path Visualization and Session Performance Evaluation with Log Files from The European Library (TEL)

Katrin Lamm, Thomas Mandl, Ralph Kölle

Institute of Information Science and Language Technology, University of Hildesheim,
Marienburger Platz 22
D-31141 Hildesheim, Germany

{lammka, mandl, koelle}@uni-hildesheim.de

Abstract

Our approach to the Log Analysis for Digital Societies (LADS) task of LogCLEF 2009 is to define three different levels of performance: success, failure and strong failure. To investigate the log files we adopt qualitative and quantitative methods. In a more qualitative approach we attempt to identify intercultural differences in user patterns by visualizing the user interactions with The European Library (TEL). On the quantitative level we calculate the relative frequencies of the different performance levels in relation with several aspects e.g. whether a user used the advanced search option or not.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.7 Digital Libraries---User issues; H.3.3 Information Search and Retrieval---Search process

General Terms

Performance, Human Factors, Experimentation, Algorithms

Keywords

Log File Analysis, Search Path Visualization, Performance Evaluation

1 Introduction

This paper describes an approach to analyze logs from The European Library (TEL) within the LogCLEF track at the Cross Language Evaluation Forum (CLEF) 2009. Many different approaches have been applied to log file analysis (Jansen et al. 2009). We intend to identify deviations between usage behaviour of users from different countries. These differences can be found in many aspects and currently it is unknown where the most relevant observation in this regard can be made. Consequently, we adopted an open approach, which allows the identification of differences in many areas and at various steps of the process. We believe that user path information could be a key to many findings.

Our approach uses the human visual capabilities to find trends and patterns by providing a visualization of user paths. This is a typical approach taken in data mining (Tan et al. 2006). We particularly decided to employ the hyperbolic tree view which provides a visualization of focus and context (Lamping et al. 1995) and which has been applied for showing large hierarchies of data (Heo & Hirtle 2001).

In section 2 we define the meaning of performance in this context followed by a detailed description of our approach of visualizing the sequence of interactions (section 3). The rest of the paper is organized as follows: Section 4 provides a brief description of the algorithm used to process the log files. Section 5 gives some introductory figures to get a first impression of the available data. Sections 6 and 7 present and analyze the results of our experiments and section 8 contains the conclusion and outlook.

2 Definition of Performance

In our approach to analyze logs from The European Library (TEL) we assume that there are indicators, which suggest that a particular session was successful or not. One such indicator is when the user chose the *Available at Library* link to view the record in a particular national library interface[1]. In our judgment this action indicates that the user came across an interesting document according to the query and hence the session was successful.
We developed a definition of performance by examining and interpreting the actions recorded in the log files. Therefore it is possible that our operational definition does not include all possible aspects of success. As this study is not meant to produce an ultimate definition of search success, but rather to try out new methodologies for log file analysis, we do not believe that this is a problem in these experiments.

Within this paper we define and evaluate the following three different levels of performance:

- **success** - a session is considered to be successful if one of the following actions is carried out at least once: available_at, see_online, option_print, option_save_reference, option_save_session_favorite, option_send_email, service
- **failure** - a session is considered to be not successful if none of the actions above is carried out during the session
- **strong failure** - a session is considered to be especially not successful if none of the actions above is carried out and the user never uses the possibility of viewing a full record (view_full)

3 Search Path Visualization

To enable a more qualitative human assessment we visualized the sequence of the individual user interactions with the interface of TEL. Therefore we generated several XML files following the GraphML file format[2]and adjusted an existing example application for a simple social network visualization[3] for the interactive data visualization.
In order to allow the analysis of multilingual search behaviour we created visualizations for German, Spanish, French, British, Italian, Dutch, Polish and US users, as they could be identified from by their IP addresses.

In each case we provide four different views[4]:

- **frequency** - the size of the edges indicates how many times the first action is followed by the second action
- **success** - the size of the edges indicates how many sessions that proceed through the second node are successful
- **failure** - the size of the edges indicates how many sessions that proceed through the second node are not successful
- **strong failure** - the size of the edges indicates how many sessions that proceed through the second node are especially not successful

For reasons of large deviations in the counters of different paths the size of the edges is scaled logarithmically.
Due to the fact that the XML files are rather large and for reasons of better clarity we do not display the complete graphs. For frequency graphs we show five levels and for the three performance graphs (success, failure and strong failure) we display ten levels. The frequency view for all users (without dividing them by country) cannot be provided, as even five levels exceed the graph size which the tool can show. The three performance graphs for all users show five levels each. Apart from that only paths where the respective counter (success, failure, strong failure) was not equal to zero are displayed.

[1] See http://www.uni-hildesheim.de/logclef/ for a more detailed description of the individual user interactions recorded in the log files.
[2] http://graphml.graphdrawing.org/
[3]The application is available at http://flare.prefuse.org/download
[4]The interactive visualizations can be found at http://app01.iw.uni-hildesheim.de/logclef/visualizations.zip

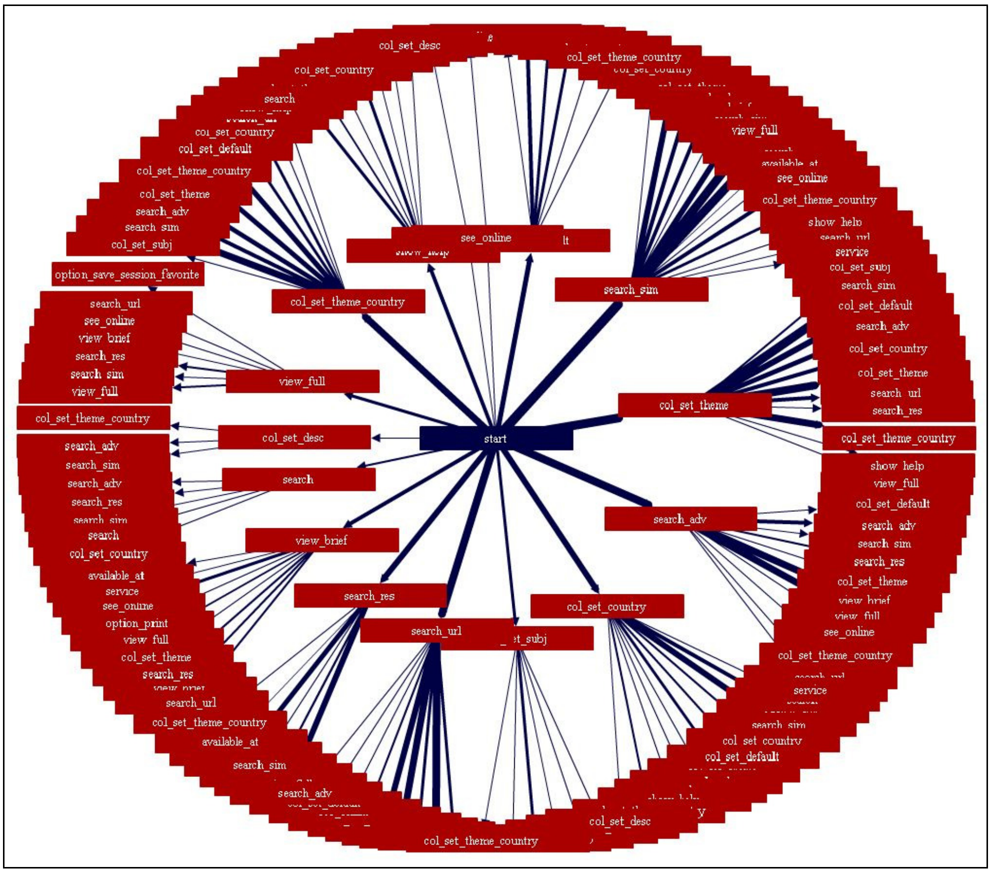


Figure 1. Success view (all users) – start configuration.

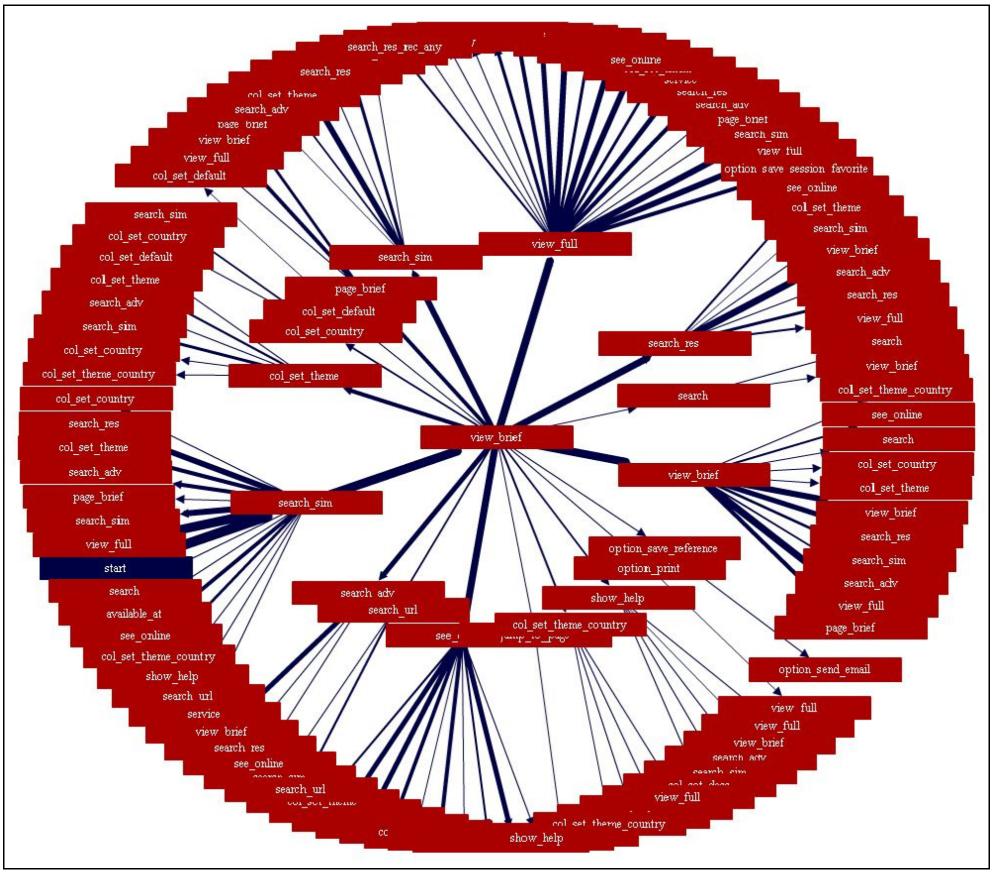


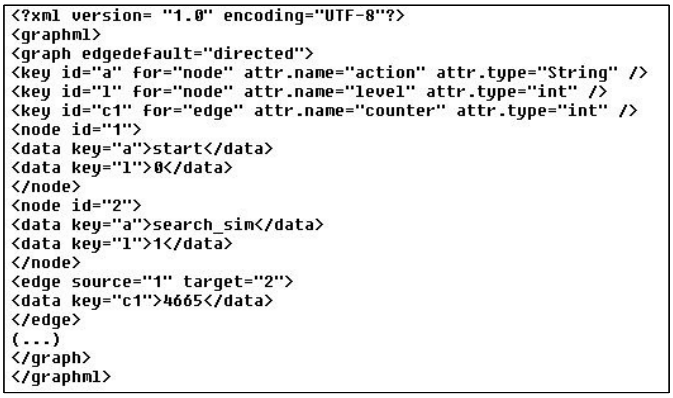
Figure 2. Success view (all users) – search path navigation.

Figure 1 shows a screenshot of the success view for the group of all users. The graph begins with a common *start* node, from which edges connect to all possible first interactions with TEL. In the interactive version the user behaviour can be retraced by navigating through the different search paths. Clicking on the next action (e.g. search_sim) causes the graph to rotate and displays this action in the middle of the screen.
Figure 2 shows another screenshot of the success view for the group of all users. This time the graph has been rotated and the action *view_brief* is displayed in the middle of the screen. View_brief stands for viewing the short title list of records and is one of the possible second interactions with TEL after users started their search from the simple search form (search_sim).

4 Algorithm for Log File Analysis

We developed a log file analyzing algorithm and applied it to the provided search action log file. The algorithm is implemented in Java using JDBC as API to access the PostgreSQL database containing the log data. In the following we briefly describe our course of action:

For reasons of database normalization we first created two additional tables. In the first table the start- and endtime for each individual session is collected. The second table contains all interactions of these sessions. Session identification has been implemented via session identifiers. Logs with the same session ID were regarded as being part of the same session. Since not the whole database could be treated at once, we sampled the log file in 48h intervals. To prevent sessions from being cut in two at the intersection of two intervals we incorporated a 10h overlap for sessions ending in the next interval. Therefore the cut-off length for sessions varies from 10h if the session starts near an intersection of two intervals to 58h if the session starts at the beginning of an interval. However, the chance that a session exceeds 10h is rather small.
Before starting the log analysis data cleaning was performed in order to ease and enhance data analysis. In this regard sessions with only one system interaction were excluded from the database, as the typical search consists at least of two interactions: search and sift through the results. Sessions with missing information on the performed user interaction (missing value) were excluded as well. The original records include several actions that stand for the user calling up the help function (e.g. show_help_help/german/search_s). For simplification of further analysis we decided to combine these actions in the action *show_help*. The same applies to actions that stand for using the *Link to other services* link (e.g. service_netherlands). Using this link, we assume, the TEL user gets the possibility to search for the respective record in services like bookshops (e.g. Amazon), search engines (e.g. Google) or other web services (e.g. Wikipedia). These actions were combined in the action *service*. Data cleaning eliminated 1.93 % of the data entries (not sessions) from the original action log file.
Subsequently, the performance of each session according to the definition of performance (see section 2) was determined.
For country recognition we integrated the *IP-to-Country Database*[5] (last updated on June 03 2009). As the last two octets were missing in the logged IP addresses it was not possible to locate all users (24.11% of the sessions could not be clearly assigned).
In the next step we generated the XML files needed for the search path visualization. An example of such an XML graph is shown in figure 3.



```
<?xml version= "1.0" encoding="UTF-8"?>
<graphml>
<graph edgedefault="directed">
<key id="a" for="node" attr.name="action" attr.type="String" />
<key id="l" for="node" attr.name="level" attr.type="int" />
<key id="c1" for="edge" attr.name="counter" attr.type="int" />
<node id="1">
<data key="a">start</data>
<data key="l">0</data>
</node>
<node id="2">
<data key="a">search_sim</data>
<data key="l">1</data>
</node>
<edge source="1" target="2">
<data key="c1">4665</data>
</edge>
(...)
</graph>
</graphml>
```

Figure 3. XML graph.

[5] The IP-to-Country Database is available at http://ip-to-country.webhosting.info/

The nodes contain the id of the node, the performed action and the level within the graph. The edges record the source and target node id and the counters for frequency, success, failure or strong failure. To construct the XML graphs we first collected the required information in nine additional tables, one per language and one table including the search path information for all users. This data structure was the input for the generation of four XML graphs for the four different views described in section 3.

5 Descriptive statistics

In table 1 we provide some descriptive statistics on the occurrence of sessions within the three levels of performance (see section 2).

Table 1. Performance statistics.

	total		**success**		**failure**		**str. failure**	
	abs.	%	abs.	%	abs.	%	abs.	%
all	191781	100	24937	100	107545	100	59299	100
de	9405	4.90	1118	4.48	5341	4.97	2946	4.97
es	12105	6.31	1643	6.59	7282	6.77	3180	5.36
fr	12574	6.56	2023	8.11	6491	6.04	4060	6.85
gb	7249	3.78	872	3.50	4299	4.00	2078	3.50
it	11979	6,25	1505	6.04	6885	6.40	3589	6.05
nl	6171	3.22	1287	5.16	3165	2.94	1719	2.90
pl	7719	4.02	924	3.71	3632	3.38	3163	5.33
us	14953	7.80	1828	7.33	8648	8.04	4477	7.55
average	10269,38	5.35	1400	5.61	5717,88	5.32	3151.50	5.31

The percentages express the country's fraction of the whole dataset. For example 1118 successful German sessions represent 4.48% of all successful sessions (24937). It can be seen from table 1 that the percentages do not vary much between the three levels of performance, which suggests that there may not be a lot distinguishable differences between the different user groups. The column entitled *total* refers to the number of sessions identified and *average* corresponds to the countries considered.

6 Analysis of User Path Information

As already pointed out in section 3 the interactive visualizations can be found and explored online. In this section we present four different search paths: the most frequently used search path (frequency), the most frequently used successful search path (success), the most frequently used unsuccessful search path (failure) and the most frequently used especially unsuccessful search path (str. failure). Table 2 illustrates these different search paths for the first ten interactions.

Since we could not detect striking differences during the qualitative analysis of the user behaviour for different countries we only show the search paths for the group of all users. The success column contains missing values as after level 7 the most successful search path was ambiguous.

At this point we want to report some observations. Firstly the most frequently used and the most unsuccessful search path are identically for six of the nine user groups (British, Dutch and Polish users deviate). If we do not assume that most of the sessions (see also table 1) and the most frequently used search path are not successful, we might have to rethink our operational definition of failure. Maybe there are users that are already satisfied by having the possibility to view a full record (view_full). Maybe some TEL users use the library primarily for informative reasons.

As to the most frequently used search path we have three different search patterns. British, Dutch, Italian and Spanish users act like we have seen before in the case of all users. They submit a query and then view the results.

German, French and US users submit a second query after viewing two full records and a third query after viewing again two full records. A possible interpretation of these differences is that we deal with two types of users here. The first type prefers to sift through the list of search results; he submits his query and then examines at least eight documents without rephrasing his query once. The second type, however, prefers rephrasing his queries subsequently; he only examines a few documents of the result list in detail before rephrasing his query. As we do not know what results the users viewed it stays open whether the first type submits more eloquent queries that return better results or whether the second type is more aware of relevant documents according to his query.

Table 2. Search paths.

level	frequency	success	failure	str. failure
1	search_sim	search_sim	search_sim	search_sim
2	view_full	view_full	view_full	view_brief
3	view_full	view_full	view_full	search_sim
4	view_full	available_at	view_full	view_brief
5	view_full	search_sim	view_full	search_sim
6	view_full	view_full	view_full	view_brief
7	view_full	view_full	view_full	search_sim
8	view_full	-	view_full	view_brief
9	view_full	-	view_full	search_sim
10	view_full	-	view_full	view_brief

Polish users differ the most from the other individually considered countries concerning the most frequently used search path. They use the short title list of records (view_brief) in order to view the results and only once view a full record (forth action). Otherwise Polish users seem to belong to the first type of users that does not rephrase his queries much. Nevertheless we compared if Polish users in fact do view_brief more often than users from other countries. Table 3 depicts the absolute frequencies and the percentage fraction of users carrying out the actions view_brief vs. view_full. In this case the percentages express the action's fraction of all actions (total) carried out by the respective user group where the user viewed results (view_brief + view_full). As can be seen from table 3 indeed Polish users do not view the short title list more often than other countries and moreover perform view_brief less often than view_full (34.1% vs. 65.9%).

Table 3. View_brief vs. view_full.

	total	view_brief		view_full	
	abs.	abs.	%	abs.	%
all	1016665	458118	45.06	558547	54.94
de	43944	24454	55.65	19490	44.35
es	65706	26286	40.01	39420	59.99
fr	54433	35024	64.34	19409	35.66
gb	51254	14721	28.72	36533	71.28
it	75099	30038	40.00	45061	60.00
nl	40709	18102	44.47	22607	55.53
pl	63795	21755	34.10	42040	65.90
us	51383	32607	63.46	18776	36.54

Our next observation supports the assumption that our operational failure definition might not capture the full picture. The search path in the last column of table 2 shows a sequence of interactions that one would expect if the session is considered to be not successful. Here the users enter a query, the result list does not fulfil their expectations, they enter a new query and so on. Concerning the most frequently used especially unsuccessful search path again we could spot three different search patterns. British, German and US users act exactly like we have seen before in the case of all users. Dutch users after the first search from the simple search form (search_sim) start their further searches from the search form in the results page (search_res), but otherwise also alternately submit a query and do view_brief. We also checked if Dutch users in fact more often perform search_res than other users, but we did not spot any differences. In this context we are not able to give an account of the behaviour of French, Italian, Polish and Spanish users, as they chose a collection from the theme list (col_set_theme) ten times in a row (for Spanish users the search path was ambiguous after level 5). We suspect that this might be due to some technical reason, because it does not make sense to change the collection more than once before any search was carried out.

The last observation refers to the most frequently used successful search path. In this context it is interesting that for all countries but Italy the fourth action is clicking the *Available at Library* link to view a record in the respective native national library interface. We have not yet found an explanation for this effect, but further research may take this into consideration. For Italian users the most frequently used successful search path is identical with the most frequently used search path, which might be a further reference to Italian users belonging to the first type of users introduced above.

7 Relative Frequencies of Performance Levels

In order to get quantitative results we calculated the relative frequencies of the three levels of performance for three different aspects which are described below:

- **advanced search** - the relative frequencies of success, failure and strong failure depending on whether the user uses the advanced search form (search_adv) during the session or not
- **number of interactions** - the relative frequencies of success, failure and strong failure depending on the number of interactions with the system
- **session duration** - the relative frequencies of success, failure and strong failure depending on the duration of the sessions (calculated for intervals of 5 min.)

Except of course for the session duration condition where the relative frequencies are calculated for intervals of five minutes the relative frequencies are calculated for the number of interactions with the system.

Number of Interactions

In a preliminary step to determine patterns of search performance, we investigated if there is a link between the number of interactions with the system during one session and the level of performance. To address this issue we compared the performance levels for sessions with 2, 3, 4 etc. interactions. We calculated the relative frequencies by dividing the number of sessions (e.g. successful sessions) by the total number of sessions (with e.g. 4 interactions).

Figure 4 shows these relative frequencies for the first 54 interactions of all users. The first 54 interactions were used as cut-off level, because of statistical reasons. Within this interval, the number of sessions per number of interactions amounts to at least 100 sessions.

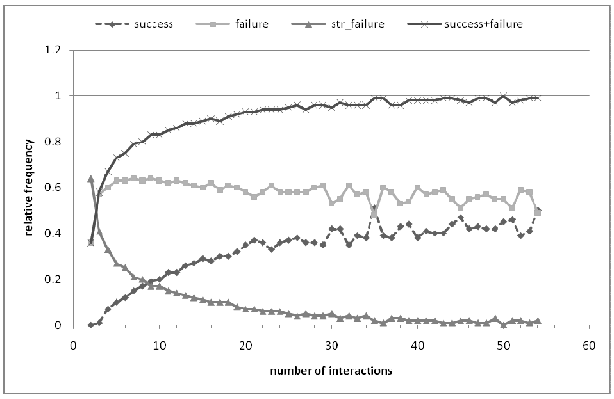


Figure 4. Relative frequencies in relation to number of interactions (all users).

If we look at figure 4, perhaps not surprisingly, we see that the success curve grows as the number of interactions increases and the strong failure curve declines. Surprisingly, the failure curve stays almost constant at 0.6%. This effect again points to the fact that our operational failure definition might not capture the full picture. As noted earlier some of the 0.6% TEL users might indeed be satisfied by using the library primarily for informative reasons, e.g. by reading the full records. This point of view is also supported by the fact, that this fraction remains almost constant even for 40 and more interactions. In our opinion this long session durations indicate successful sessions, because otherwise the users would have abandoned their search earlier. That is why we included a forth curve to this diagram where we added up the values for success and failure. The curve tends to one with increasing session length. This leads to two possible interpretations. Either people that search long enough will finally be successful or people, who are successful, will search longer. If and which explanation is the right one, would have to be decided in another user test. With respect to the different countries there are no distinguishable differences.

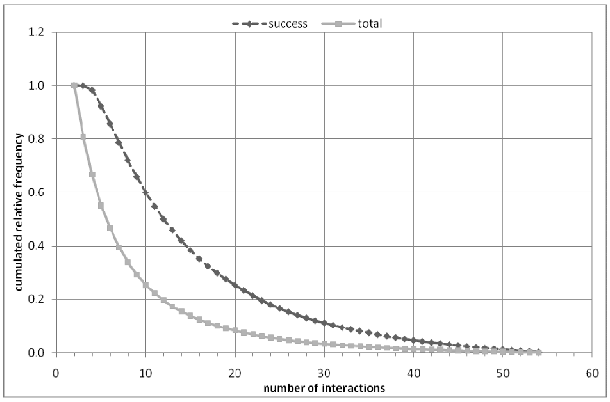


Figure 5. Cumulated relative frequencies in relation to number of interactions (all users).

Figure 5 shows the cumulated relative frequencies of all users for successful and for all sessions (total). Again the calculation is cut off after the first 54 sessions. The cumulated relative frequencies are calculated by dividing the number of sessions per number of interactions by the total number of sessions. We decided to present this diagram, because it shows illustrative that 0.6% of the successful sessions have 10 or more interactions and that only 0.25% of all sessions have 10 or more interactions.

Advanced Search

Considering all the sessions 13.39 % of all search actions were started from the advanced search form. The trend of the relative frequencies of the three levels of performance for users that use and users that do not use the advanced search form (search_adv) is qualitatively consistent with the results for the number of interactions condition. Hence we decided only to show a comparison of the relative frequencies of the successful sessions. As already described above, the relative frequencies are calculated by dividing the number of successful sessions (with or without using the advanced search) by the total number of sessions.

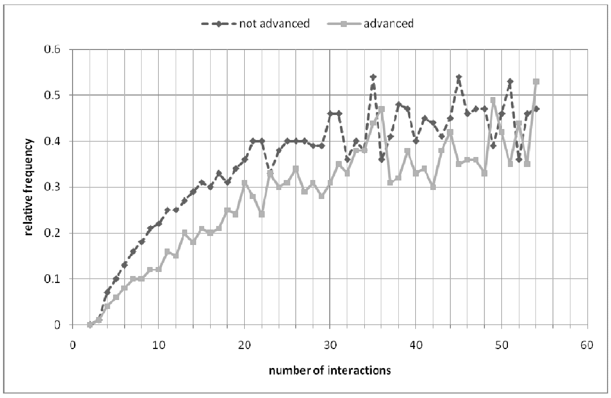


Figure 6. Relative frequencies for advanced search (all users).

Figure 6 shows these relative frequencies for the first 54 interactions of all users. This diagram shows two things. Firstly, users searching longer are rather willing to use search_adv and secondly in the interval between 5 and 20 interactions more users are successful not using the advanced search. This trend also appears for the individually considered countries.

Session Duration

Figure 7 compares the levels of performance for sessions with 5, 10, 15 etc. minutes. The calculation of the relative frequencies corresponds to the calculation described in matters of the number of interactions condition. Figure 7 reflects the relative frequencies for all users and again the cut-off level was chosen so that the number of sessions per timestamp amounts to at least 100 sessions.

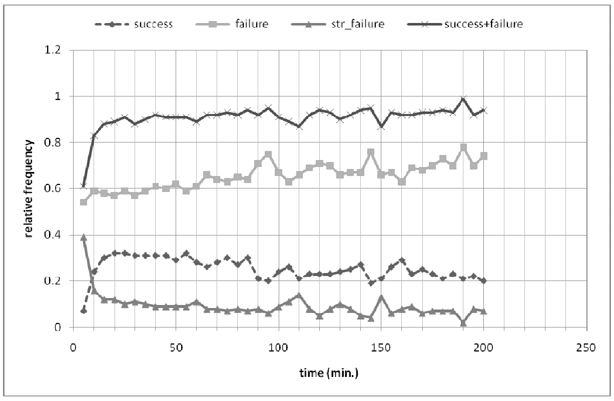


Figure 7. Relative frequencies for session duration (all users).

As can be seen from this diagram at each time (after the first 20 min.) the proportions of success, failure, strong failure and success plus failure are approximately the same. In other words, after the first 20 minutes the probability of a successful search becomes independent of the session duration. This could reflect the differences in the search speed of different users.

8 Conclusion and Outlook

The aim of this study was to experiment with new methods for log files analysis. As a starting point we developed an operational definition of search performance with three different levels of performance. To enable a more qualitative human assessment we visualized the sequence of the individual user interactions with the interface of TEL. Both, the more qualitative analysis of the search path visualizations as well as the more quantitative analysis of the log files have shown inconsistencies within the data which suggest that our operational definition of performance has to be modified in the following way:

- **success** - a session is considered to be successful if one of the following actions is carried out at least once: available_at, see_online, option_print, option_save_reference, option_save_session_favorite, option_send_email, service, *view_full*
- **failure** - a session is considered to be not successful if none of the actions above is carried out during the session

One proposal for an additional success indicator for future research is whether the session ends with a search or not. This could imply that the user could not find the information he/she was looking for and therefore the session could be evaluated as not successful.

During the qualitative analysis of the user path information we observed some differences between users from different countries, e.g. that there seem to exist two prevailing search patterns (cf. most frequently used search path section 6). Whereas one group of users (British, Dutch, Italian, Polish and Spanish users) seems to examine more documents after the first query, another group of users (German, French and US users) seems to rephrase their queries more often. But of course further user tests are required to determine whether the spotted search patterns can serve as categories to differentiate between several types of users.

Although further research is needed to confirm our findings, basically we can say now that it is possible to investigate user performance from log files and that our refined definition of performance at least in the context of TEL users accounts for the user behaviour.

References

[1] Mandl, Thomas; Giorgio, Di Nunzio; Schulz, Julia Maria; Yeh, Alexander (2009): LogCLEF 2008: the CLEF 2009 Cross-Language Logfile Analysis Track Overview. In this volume.

[2] Jansen, Bernard; Spink, Amanda; Taksa, Isak (eds.) (2009): Handbook of Research on Web Log Analysis. Idea Group Reference: Hershey et al.

[3] Tan, Pang-Ning; Steinbach, Michael; Kumar, Vipin (2006): Introduction to Data Mining. Addison-Wesley: Boston et al.

[4] Heo, Misook; Hirtle, Stephen C. (2001): An empirical comparison of visualization tools to assist information retrieval on the web. In: Journal of the American Society for Information Science and Technology (JASIST) 52(8): pp. 666-675

[5] Lamping, J., Rao, R., and Pirolli, P. (1995). A focus+content technique based on hyperbolic geometry for viewing large hierarchies. In: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Denver. ACM