Capturing and Analyzing User Behavior in Large Digital Libraries

Giorgi Gvianishvili, Jean-Yves Le Meur, Tibor Šimko, Jérôme Caffaro,
Ludmila Marian, Samuele Kaplun, Belinda Chan, and Martin Rajman

European Organization for Nuclear Research
CERN, IT Division,
CH-1211, Geneva 23, Switzerland
{giorgi.gvianishvili,jean-yves.le.meur,
tibor.simko,jerome.caffaro,
ludmila.marian,samuele.kaplun,belinda.chan}@cern.ch
http://www.cern.ch

Swiss Federal Institute of Technology
EPFL, Artificial Intelligence Laboratory,
CH-1015, Lausanne, Switzerland
martin.rajman@epfl.ch
http://www.epfl.ch

Abstract. The size of digital libraries is increasing, making navigation and access to information more challenging. Improving the system by observing the users' activities can help at providing better services to users of very large digital libraries. In this paper we explain how the Invenio open-source software, used by the CERN Document Server (CDS) allows fine grained logging of user behavior. In the first phase, the sequence of actions performed by users of CDS is captured, while in the second phase statistical data is calculated offline. This paper explains these two steps and the results. Although the analyzed system focuses on the high energy physics literature, the process could be applicable to other scientific communities, with and international, large user base.

Keywords: Invenio, CDS, Very large digital library, Log analysis, User behavior study analysis

1 Introduction

Digital libraries are playing a strategic role in the showcasing of research done by an institution or university, since 1988. Large scientific communities rely on the digital libraries as a primary resource for storing and acquiring information. One of the challenges is to make navigation in large amounts of data as intuitive as possible. Our goal is to concentrate on the users' specific needs in order to improve and optimize access to information. In addition to the successful survey done in 2008 on the information resources in High-Energy Physics [8] the behavior of CERN Document Server (CDS) users has been studied.

CDS is an instance of the Invenio software, which is developed and maintained at CERN. The number of records in CDS exceeds 1 million and continues to grow, while the number of unique users is more than 40 000, making CDS one of the largest digital libraries in the physics domain [14]. We have logged users' interaction with the search engine and analyzed them using automated data processing techniques. This automated approach helped to reveal important patterns, which are difficult or sometimes even impossible to spot by a human, due to the large amount of data involved.

In order to understand which options users can select and which of them were used or ignored, we first describe the CDS production environment, underling its core functionalities and possibilities. We then describe the logging phase and the type of information collected. Finally, we explain the automated extraction of additional information and the results obtainedreturned.

2 System Description

Invenio [1] [2] [3] [4] is a digital library system which is freely available under GNU General Public License. Invenio consists of a set of modules for maintaining intermediate to large digital library services. It has been actively used at CERN since 2002. Besides CERN, it is also used in diverse scientific institutions and universities worldwide like EPFL [5], DESY [6] and others. The system can handle not only articles and books, but also theses, photos, videos, etc.

Records maintained by Invenio are organized in collections that can be defined on top of any query. Users are offered either simple or advanced search interfaces. They can query specific fields, such as title, author, etc., sort the results or apply a ranking criteria (like word similarity).

Users can restrict their search to a set of specific collections or sub-collections, and the results returned can be merged into a single list.

Users can also customize the output format of the results: by default a summary of the results is displayed (brief HTML) but other formats such as detailed HTML, HTML MARC and others are also provided. Invenio has been translated into 26 languages and supports Unicode for information retrieval.

Users can also register an account in order to access restricted collections or to use Web 2.0-like services (baskets, alerts, etc.).

In CDS, there are approximately 8 000 registered users, representing a large portion of the high energy physics community. However, the majority of users are not registered ($\sim$ 40 000 users).

Most of the content maintained in CDS are articles and preprints, coming mainly from the high energy physics domain.

3 User Logging

3.1 First Phase

Invenio software allows user activities to be logged in real-time into MySQL tables. Standard logs collected from the web-server do not provide sufficient information for observing users' interactions with the system. In addition to web-server logs we can store information about the query recall, the status of the user and the rank of downloaded documents. The main challenges arising in this phase are:

- Defining the data and the events to be captured
- Preserving the relationship between stored data
- Making logging transparent to end users

Two tables contain information about user queries, with the following data: user id, date, host name, IP address, HTTP referrer and query recall. Recall is stored as the list of records unique identifiers. Other tables are dedicated to log the downloads and the accesses to detailed page view information. For each record, download time, client host, user id, file format, HTTP referrer and display position are logged. The download table is used not only to store local file downloads, but also downloads of documents which are hosted on remote servers, and which cannot be extracted from the web-server logs. To preserve the relationship between stored data query id, user id and IP address are used. These identifiers can uniquely identify the history of a user's action in the system.

3.2 Second Phase

After the logs have been collected, a post-processing phase that is more computationally intensive is executed offline. Four types of counts are extracted from the logs:

- Number of detailed page views: for each record we count the occurrences of record abstract being viewed
- Number of downloads: for each record we count the occurrences of the associated full-text being downloaded
- Number of displays: for each record we count the occurrences of the record being listed on the results pages
- Number of seens: for each record we count the occurrences of record being seen. We mark all records seen from the first up to the one on which an action has been performed (download/view). For example, if a user downloads record #6 we mark all records from #1 up to #6 as seen, since those records would have most probably been seen by the user. This count provides us with an approximate result of records seen, since there is no guarantee that the user has really seen those records.

These numbers, can then be used not only for information but also for ranking and for analysis of the relationship among records. They might also suggest the reorganization of the digital library for optimizing its performance. Concerning ranking, combining these counts with other attributes like freshness, citation frequency and Hirsch index is being studied within the scope of the collaborative D-Rank [13] project at Swiss Federal Institute of Technology (EPFL) and Central European Organization for Nuclear Research (CERN). The core idea of the project is to take into consideration a user's previous interaction with the system: if some subsets of documents have been downloaded or viewed, it is assumed that their importance will be preserved; on the other hand, documents which were displayed or potentially seen, but not downloaded or viewed will be considered as less important.

4 Analysis

After analyzing more than 130 000 queries maintained on CERN Document Server (CDS), it can be observed that 73.1% of users are using the English interface which is set by default. Usage of other languages is relatively equally distributed (Figure 1).

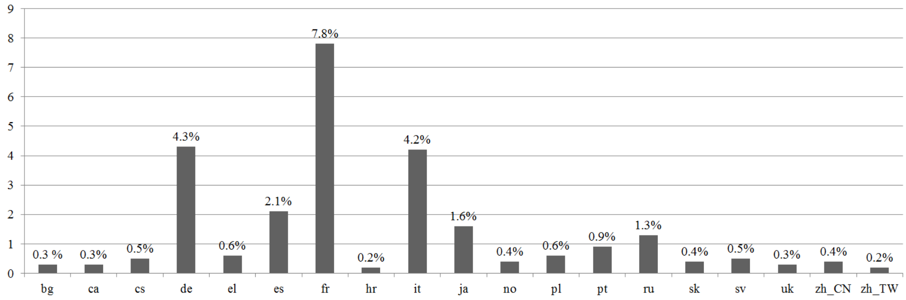

Fig. 1. Number of users using non-English search engine interface. The English interface has been used 48 507 times (73.1%).

The default ordering of the records is 'latest first', but it can be changed to display results according to word similarity criteria. As we observed, such type of adjustments are extremely rare: in the case of CDS only in 1% of the queries ranking method has been changed from latest first to word similarity. This confirms a habit that has already been observed in the past [9]. It confirms that professionals from a specific field look more for the recent publications [10]. Yet another explanation can be that options in the interface are not intuitive enough to users. Sorting has been used in less than 3% of the queries Table 1. The default setting where the descending order of results has been used is in 97% of cases.

CDS supports a wide variety of output formats, but users change the default setting (condensed display) in only 1.5% of all queries.

Advanced search was used in approximately 8% of all queries, using the default matching type most of the time (85.4%) (matching all the words). Remaining matching types and operators used are compared in Table 2 and Table 3. As we can observe from the Table 2, besides the default option users are most often using the 'Regular Expression' matching type. Such contrast of using

Table 1. Percentage of using various sorting criteria. Default ordering criteria (Latest First) was used in 97.2% of cases.

Latest First	Chronological Order	Key Title	Year	Report Number	Author	Title	Other
97.2%	1.0%	0.9%	0.3%	0.1%	0.1%	0.1%	0.3%

the default or the most advanced technique can be caused by using pre-defined queries. Operators used operators (Table 3) suggest that users prefer achieving higher precision by restricting the criteria (Google behavior). We can observe in Table 4 that the same fields are used to query CDS in bothe the simple and advanced search interfaces.

In Table 5 we can observe the 64 most often issued query terms. Although it is possible to enter Boolean expressions and years in the dedicated fields, users prefer typing them using free text query.

Table 2. Percentage of using various matching criteria, in advanced search.

Matching Type	Percentage
All of the words:	85.4
Any of the words:	1.3
Exact phrase:	1.5
Partial phrase:	0.3
Regular expression:	11.5

Table 3. Percentage of used operators in the advanced search for defining relationship among matching fields.

Operator	Percentage
AND	96.8
OR	2.3
NOT	0.9

Rank of downloads and detailed page views is shown in the Table 6. Top ranked records are downloaded/viewed on average 9 times more than ones on the 9th position. In Table 7 we can see the 10 most often displayed records with corresponding counts. The search engine returned no results in less than 1.5% of all queries. The distribution of user access through the day (Figure 2) confirms that CDS is the institutional repository.

Table 4. Percentage of using different fields in the simple and advanced search interfaces.

Field	Simple Search	Advanced Search
any field	49.1	71.8
author	15.2	12.8
title	19.5	6.5
keyword	2.8	3.5
report number	3.0	2.6
year	9.8	0.6
other	0.6	2.2

Table 5. List of most often used terms in user queries for 5 weeks, with corresponding frequencies. (Typically several terms are combined to form query.)

Term	Frequency	%	Term	Frequency	%	Term	Frequency	%
lhc	2289	3.5%	magnet	237	0.4%	energy	184	0.3%
cern	1468	2.2%	neutrino	235	0.4%	collision	184	0.3%
atlas	1324	2.0%	lhcb	234	0.4%	school	178	0.3%
physics	1023	1.5%	programme	233	0.4%	control	176	0.3%
higgs	469	0.7%	collaboration	220	0.3%	model	171	0.3%
particle	412	0.6%	trigger	218	0.3%	technical	171	0.3%
detector	404	0.6%	performance	213	0.3%	first	169	0.2%
alice	346	0.5%	quantum	213	0.3%	training	168	0.2%
data	337	0.5%	hadron	199	0.3%	electron	163	0.2%
beam	319	0.5%	bulletin	198	0.3%	tunnel	161	0.2%
lecture	299	0.5%	computing	194	0.3%	field	160	0.2%
design	287	0.4%	system	192	0.3%	experiment	158	0.2%
accelerator	284	0.4%	student	189	0.3%	academic	157	0.2%
muon	258	0.4%	collider	189	0.3%	logo	154	0.2%
theory	239	0.4%	introduction	187	0.3%	collisions	153	0.2%
calorimeter	237	0.4%	reconstruction	185	0.3%	john	147	0.2%

5 Conclusion

Thanks to its rich mechanisms Invenio is giving a lot of possibilities for observing how users are interacting with the system. The number of users and records maintained by CDS makes it one of the largest open access digital repository in science. Capturing and analyzing user logs can provide us with hints on how to improve the usability of the system. Log analysis results can be combined with other types of user behavior studies, for better understanding the user needs.

Log analysis procedure in CDS is done in two phases. In the first step logs are collected online. The second phase is run offline, for extracting detailed statistics.

Collected data can be applied to the new ranking algorithm, building recommendation systems or identifying user communities with common interests. Other possible applications are: construction of query expansion mechanisms, user interface optimization, identifying most requested queries for their opti-

Table 6. Rank on which records have been downloaded or detail page viewed, with corresponding counts. Mostly 'latest first' ordering has been used.

Rank of Results List	Download Frequency	Rank of Results List	Page View Frequency
1	1428	1	1885
2	566	2	973
3	353	3	768
4	287	4	618
5	203	5	494
6	180	6	359
7	143	7	381
8	128	8	261
9	117	9	297
≥10	4175	≥10	6676

Table 7. Top 10 most often displayed records, with corresponding seen, download and abstract view counts.

	Displays	Seens	Views	Downloads
1	237	17	22	198
2	247	23	10	130
3	358	54	24	100
4	182	9	10	97
5	139	4	5	80
6	154	9	36	76
7	238	25	26	75
8	234	17	15	63
9	106	6	0	63
10	76	4	2	58

mization, defining the most suitable time for running computationally intensive tasks and many others.

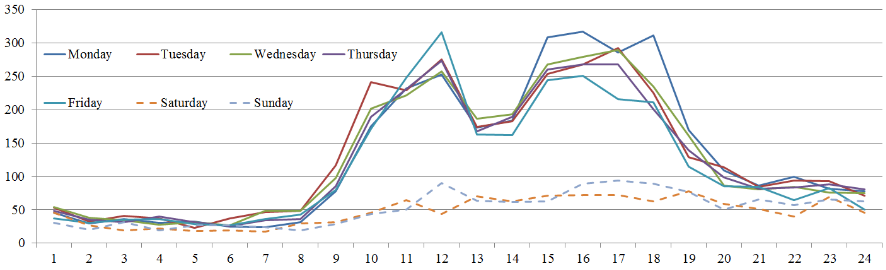

Fig. 2. Average user access through the day. By dashed lines there are denoted weekends, while with the solid line workdays.

References

1. Invenio software website, `http://invenio-software.org`
2. CERN Document Server, `http://cds.cern.ch`
3. Pepe A., Le Meur J.-Y., Simko. T.: Dissemination of scientific results in High Energy Physics: the CERN Document Server vision, (2006)
4. Pepe A., Baron T., Gracco M., Le Meur J.-Y., Robinson N., Simko T., Vesely M.: CERN Document Server Software: the integrated digital library, (2005)
5. Swiss Federal Institute of Technology, Knoweledge and Information Services `http://infoscience.epfl.ch/`
6. DESY, A Research Center of the Helmholtz Association `http://desy.de/`
7. SLAC, Stanford Linear Accelerator Center Library/SPIRES, Stanford University `http://www.slac.stanford.edu/spires/`
8. Gentil-Beccot A., Mele S., Holtkamp A., O'Connell H. B., Brooks T.C.: Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course, (2008)
9. Jones T., Cunningham S.J., Mcnab R., Boddie S. : A Transaction Log Analysis of a Digital Library , Department of Computer Science, University of Waikato, International Journal on Digital Libraries, pp. 152–169 (1999)
10. Tenopir C.: Use and users of electronic library resources: An overview and analysis of recent research studies. Washington, DC: Council on Library and Information Resources, (2003)
11. Covey D. T.: Usage and usability assessment: Library practices and concerns. Washington, DC: Council on Library and Information Resources, (2002)
12. Papatheodorou C., Kapidaki S., Sfakakis M., Vassiliou A. : Mining User Communities in Digital Libraries, (2003)
13. Vesely M., Rajman M., Le Meur J.-Y., Using Bibliographic Knowledge for Ranking in Scientific Publication Databases, Published in IOS Press, (2008)
14. Ranking Web of Worl Repositories `http://repositories.webometrics.info/top800_rep_inst.asp`