User Strategies in Query Formulations in the Internet Library Catalogue

Kazimierz Choroś
Institute of Applied Informatics, Wrocław University of Technology, Wyb. S. Wyspiańskiego 27, 50-370 Wrocław, Poland
Kazimierz.Choros@pwr.wroc.pl

Justyna Kowalska
<dronka@poczta.fm>

Izydor Statkiewicz
Main Library and Scientific Information Centre,
Wrocław University of Technology, Wyb. S. Wyspiańskiego 27,
50-370 Wrocław, Poland
Izydor.Statkiewicz@pwr.wroc.pl

Abstract

The Internet catalogue is a very easy and comfortable way of retrieving adequate literature and a practical technique for distance booking for interesting items in the library. The Main Library and Scientific Information Centre of the Wrocław University of Technology has used the electronic catalogues since 1978 and since 1995 they have also been available in the Internet network. In this paper the transaction logs containing 623,875 queries collected over 7 years have been analysed, and then the user sessions have been recognized. The article also discusses two main user strategies: expansion strategy and contraction strategy in his behaviour and in the query refinement in the Internet library catalogues.

1. Introduction

The electronic library catalogues and electronic documents become the most important information tools the library offers. The library catalogue is an index describing books, journals, reports etc. stored in the library. The electronic library catalogues totally replaced traditional paper catalogues. The interfaces of electronic catalogues are most often user-friendly, so in general the user has no need to learn how to use them. The information retrieval languages of these catalogues are very simple and intuitive, frequently offering the pre-formulated search sheets helping the definition of the query reflecting user information needs. The most common information retrieval languages used in many retrieval

systems and in Web search engines are generally based on the Boolean expressions of index terms using the AND and OR operators for conjunctive and disjunctive queries, respectively, and using the NOT operator for negation [Jansen 1998, Jansen 2000a, Jansen 2000b, Silverstein 1999, Spink 2002]. To help the user to define a query in a form of such a Boolean expression the systems also offer advanced retrieval methods based on the sheets enabling the user to express his information needs even unconsciously by logical formula.

The Wrocław University of Technology employs more than 2,000 academics. About 32,000 students are studying at 12 faculties. For many years the catalogues of the Main Library and Scientific Information Centre of the Wrocław University of Technology have been accessible for academics and all students not only in the local system in the lending library offices but also in the Internet networks. The analysis of the user queries submitted to the computer catalogues were possible due to the automatic procedure storing all Internet transactions in library catalogues in information logs reflecting the user behaviour in the system.

Such analyses have recently been performed mainly for well-known Web search engines [Park 2005, Silverstein 1999] and they have shown the user behaviour as well as troubles in formulating queries adequate for user information needs. The problem is also important in the case of digital library catalogues [France 1999, Jones 2000, Poo 2000, Ryan 2003, Warren 2001]

2. Search queries

The APIN/UDOS system [Klesta 2000] managed six library catalogues: the catalogues of books, scientific papers, journals, fiction books, foreign periodicals in Wrocław and Opole Libraries, and electronic documents. It started to operate in November 1995 and then during next 7 years all transactions in the catalogue were registered. Each transaction records the query formulated by the given user, i.e. a Boolean expression, exactly as entered by the user, the type of a search sheet used to formulate a given query, the ID of the accessed catalogue, the number of retrieved items, the indicators of item descriptions examined by the user, then which items have been booked, whether help pages have been consulted, the day and the time measured in hours, minutes, and seconds, and finally the IP number of the user computer.

Table 1. Number of queries per year.

Year	Number of queries
1996	15,230
1997	24,438
1998	38,500
1999	52,303
2000	79,393
2001	154,138

Year	Number of queries
2002	259,873
Totally	623,875

In 2003 a new system was installed, continuing the functionality and usability of electronic catalogues. Such a transition of catalogue data is necessary in libraries even in case of the change of the computer systems [Browne 1996]. It was the reason that the transaction logs from the years 1996-2002 were analysed enabling us to make some comparative analysis. Table 1 and Figure 1 present the dynamic growth of the number of user queries submitted to the library catalogues in the Internet network.

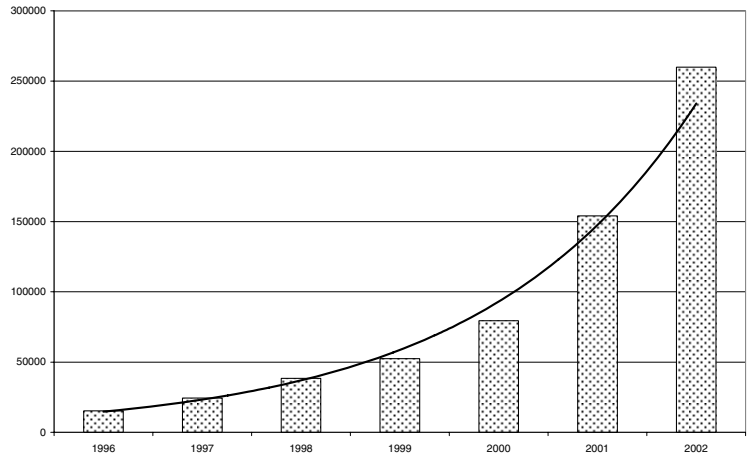

Figure 1. Growing tendency of the number of queries.

Table 2 presents the number of queries using Boolean operators. We found that the use of the negation operator is extremely rare. This operator, which in general is the most difficult to design and to implement in a real system is practically not used. It may be surprising that the Boolean expressions with parentheses are used, so we cannot state that users do not formulate complex queries. The use of more complex queries with parentheses is not very important. It confirms the observations made for other systems [Jansen 1998, Jansen 2000b, Spink 2002]. It also leads to the conclusion that may be users, for example new students, should be better trained in using computer catalogues in the library. The investigations of Topi and Lucas [Topi 2005] have shown that both Boolean training and the use of an assisted interface improved the user search performance.

The queries to the principle catalogue of books were almost only in a simple form, i.e. only simple terms (or even one term) in the default conjunction relation. The advanced form of a query - potentially leading to a better formulation of the query -was practically not used.

Table 2. Number of queries using Boolean operators per year.

Year	AND		OR		NOT		Any operator and parentheses	
	Number	[%]	Number	[%]	Number	[%]	Number	[%]
1996	7,863	51.63	301	1.98	22	0.14	275	1.81
1997	8,172	33.44	622	2.55	122	0.50	289	1.18
1998	11,705	30.40	430	1.12	96	0.25	189	0.49
1999	16,709	31.95	687	1.31	193	0.37	287	0.55
2000	13,916	17.53	4,575	5.76	558	0.70	1,005	1.27
2001	67,449	43.76	4,267	2.77	314	0.20	25,424	16.49
2002	137,385	52.87	1,982	0.76	130	0.05	18,233	7.02

Figure 2. Retrieval sheet for a simple query of the interface of the book catalogue. The user could formulate a simple query as the conjunction of the following criteria: author name, first name, terms of title or keywords, library call number, year of publication, and language.

The next table, Table 3, shows that the greatest number of advanced queries was submitted to the catalogue of scientific papers of the academics of the Wrocław University of Technology. About 15% of queries concern the ID number of a given academic, so these queries were not in the form of Boolean expressions. Similarly also the queries submitted by external search engines were not analysed.

Table 3. Number of simple queries and advanced queries in each catalogue.

Catalogue	Simple Queries	[%]	Advanced Queries	[%]
Books	417,731	96.6	8,622	2.0
Scientific papers WUT	93,622	61.0	37,932	24.7
Journals	75,559	86.3	10,332	11.8
Fiction books	26,540	88.7	3,085	10.3
Foreign Periodicals in Wrocław and Opole Libraries	25,281	86.5	3,049	10.4
Electronic documents	11,971	81.8	1,733	11.8

3. User Sessions

A user session is generally defined [He 2002, Silverstein 1999] as a sequence of queries of a single user formulated in one search. Sessions have been recognized as a series of queries sent from the same computer, identified by the computer IP number and sent in a given time interval. We used only one criterion, time criterion to recognize the sessions, we did not analyse the change of terms used in a series of queries, although we defined such a situation as a kind of user strategy in a query refinement process.

To recognize the characteristics of the user sessions the queries presented in 2001 to the library catalogue have been examined. For every query the time between the query and the previous one was calculated if both queries were sent from the same computer. These two queries were treated as two queries of the same session if this time did not exceed a given threshold. The total time of a session has been then calculated, as well as the number of queries in one session. In the set of queries from 2001 only 6,175 sessions were recognized because the additional restriction was taken into account, the session could not last more than 3 hours. It reduced the number of session by half. The majority of long sessions were the sessions of a library service. The library workers also accessed the library catalogue in the Internet network for their internal librarian purposes and their connections lasted generally more than 10 hours a day, because the library was functioning from 8 am. to 8 pm.

The average time of a single session was 10,7 minutes. Most of them were rather short, less than 5 minutes. The diagram in Fig. 2 presents the distribution

of time of the user sessions. The diagram shows that the majority of the sessions were very short, 1-2 minutes, or even less than one minute. In such a short time a user is able to formulate only one short query, rather simple query expressed by a single term and then to book an item. To reformulate a query or to examine the search result much more time is needed.

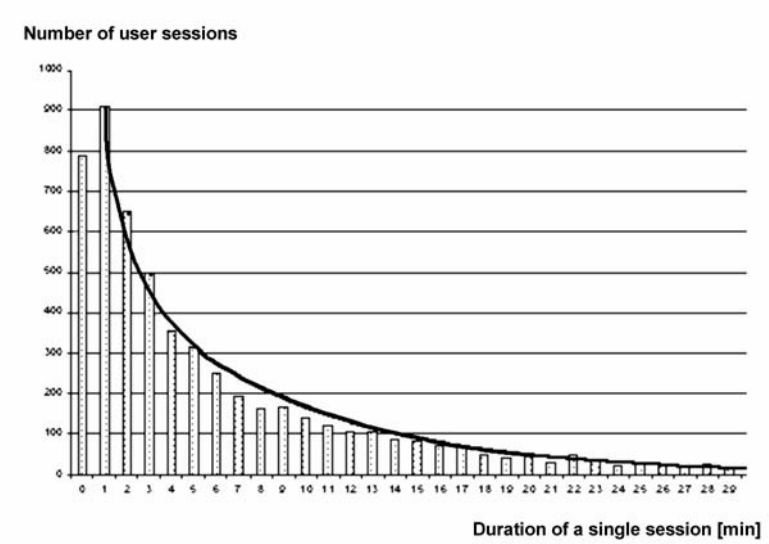

Figure 3. The distribution of the time of a single user session.

The period of 10 minutes seems to be enough to make a search, examine the results and in consequence to book a retrieved item. Such sessions of about 10 minutes are 74,1% of the whole number of recognized sessions. Only 17,5% of sessions lasted longer than 15 minutes (and less than 3 hours). The average number of queries in a single session was 3,76. So, the average user formulates more than one query in one session. It raises a question what the strategy of formulating this sequence is.

Table 4. Number of library catalogues accessed in one session.

Number of Library Catalogues	Number of Sessions	[%]
1	5,014	83.93
2	652	10.91
3	214	3.58
4	63	1.05
5	25	0.42
6	6	0.10

Only in 16% of sessions we observed queries submitted to more than one catalogue. The conclusion is that some users tried to improve their results by consulting several library catalogues.

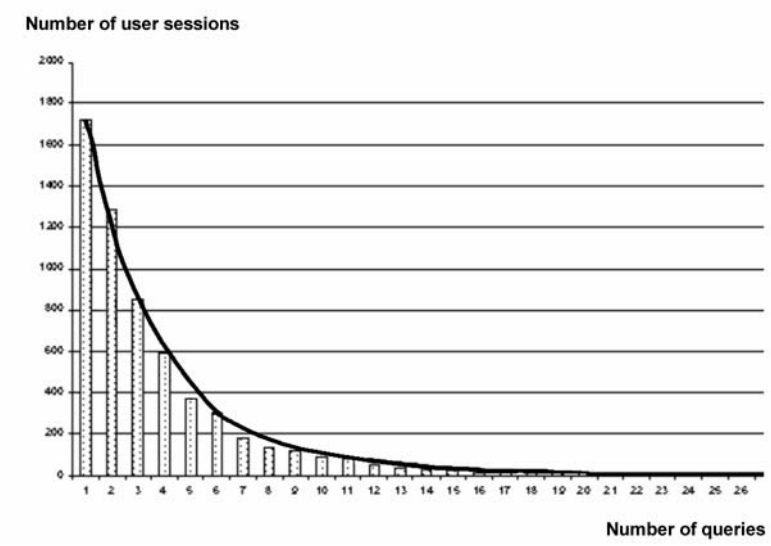

Figure 4. The distribution of the number of queries in a single user session.

4. User Strategies

Two strategies can be defined in a user behaviour: contraction strategy, i.e. the strategy leading to a more specific query, and expansion strategy, i.e. the strategy leading to a more general query. An expansion strategy, a rather rare strategy, is generally achieved by adding a new term to a disjunctive query. On the other hand, the contraction strategy, which was much more frequently undertaken was observed when the user received too many items in a system response, in which case to limit the result number, the user adds a new search term to the query. But in this case the added term was in conjunction with previous terms. Unfortunately in many cases it led to a zero response. The conjunction is a very restrictive Boolean operator.

The set of 6,175 recognized sessions included 1,007 contraction sessions and only 83 expansion sessions. It means that the first user query is much more frequently too general than too specific. The most frequent user action was adding a new term with AND (*) or OR (+) operators.

Examples from the transaction logs of the different subtypes of two main user strategies:

I. CONTRACTION STRATEGY
- adding a new term to the conjunctive query
  Example 1:
    the originally formulated query: IEEE*communications
    and the modified query: IEEE*communications*transactions
- replacement of one term by another, more specific
  Example 2:
    AN:Kowalski*AI:Jan*(mieszanki*mineralno*asfaltowe)
    AN:Kowalski*AI:Jan*(mieszanki*mineralne*asfaltobetonowe)

II. EXPANSION STRATEGY
- adding a new term to the disjunctive query
  Example 3:
    koszty*paliwa
    (koszty+eksploatacja)*paliwa
- replacement of one term by another, more general
  Example 4:
    AN:Nowak*AI:Piotr*plane
    AN:Nowak*AI:Piotr*transport

The other observed actions of the users are:
- adding the category (an - author name in the example) of a search term
  Example 5:
    Nowak*Piotr
    an:Nowak*Piotr
- auto correction of an spelling error
  Example 6:
    iinformation*retrieval*system
    information*retrieval*system

5. Conclusions

The analyses of transaction logs of library catalogues show that the end user expects simple retrieval mechanisms applied in the catalogue. The user prefers to formulate a rather simple query, although, the user being not satisfied by the system response is capable to continue his search by reformulating his initial query several times. It may take even several minutes. It seams to be reasonable to maintain several specific library catalogues, as it has been done in the Library of Wrocław University of Technology, because we observe the differences in the complexity of user queries submitted to the different catalogues. It seems that more training is necessary for users in Boolean query formulations because not

all the opportunities of the computer catalogues are applied. For example, the negation operator is practically not used, although its usefulness in a query formulation cannot be denied.

The dominant role of a conjunction operator in user queries does not mean that this operator is well-known and preferred by the users. Rather, it is due to the application of pre-formulated retrieval sheets. The criteria proposed on these sheets are mostly in conjunction relations.

The information about the user behaviour gathered in the transaction system logs can be very useful in managing the catalogues. The user sessions do not generally exceed 15 minutes, so we should take into account the frequency of their access and this maximum time of a single session in planning the accessibility to the network catalogue for a given user population.

6. References

[Browne 1996] S.V. Browne, J.W. Moore, "Reuse library interoperability and World Wide Web", 1996, http://www.netlib.org/srwn/srwn20.ps

[France 1999] R.K. France, N.L. Terry, E.A. Fox, R.A. Saad, J. Zhao, "Use and usability in a digital library search system", 1999, http://www.dlib.vt.edu/Papers/Use_usability.PDF

[He 2002] D. He, A. Göker, D.J. Harper, "Combining evidence for automatic Web session identification", Information Processing and Management, 2002, 38, pp. 727-742.

[Jansen 1998] B.J. Jansen, A. Spink, J. Bateman, T. Saracevic, "Real life information retrieval: A study of user queries on the Web", SIGIR Forum, 1998, 32(1), pp. 5-17.

[Jansen 2000a] B.J. Jansen, "An investigation into the use of simple queries on Web IR systems", Information Research: An Electronic Journal, 2000, 6(1), http://jimjansen.tripod.com/academic/pubs/ir2000/ir2000.pdf

[Jansen 2000b] B.J. Jansen, A. Spink, T. Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the Web", Information Processing and Management 2000, 36, pp. 207-227.

[Jones 2000] S. Jones, S.J. Cunningham, R. McNab, S. Boddie, "A transaction log analysis of a digital library", International Journal on Digital Libraries, 2000, 3, pp. 152-169.

[Klesta 2000] D. Klesta, I. Statkiewicz, „Komputerowy zintegrowany system biblioteczny APIN", Elektroniczny Biuletyn Informacyjny Bibliotekarzy EBIB 2000, 10, (in Polish) http://www.oss.wroc.pl/biuletyn/ebib10/apin.html

[Park 2005] S. Park, J.H. Lee, H.J. Bae, "End user searching: A Web log analysis of NAVER, a Korean search engine", Library & Information Science Research, 2005, 27, pp. 203-221.

[Poo 2000] D.C.C. Poo, Toh Tech-Kang, C.S.G. Khoo, "Enhancing online catalog searches with an electronic reference", The Journal of Systems and Software, 2000, 55, pp. 203-219.

[Ryan 2003] S.M. Ryan, "Library Web site administration: A strategic planning model for the smaller academic library", The Journal of Academic Librarianship, 2003, 29(4), pp. 207-218.

[Silverstein 1999] C. Silverstein, H. Marais, M. Henzinger, M. Moricz, "Analysis of a very large Web search engine query log", SIGIR Forum, 1999, 33(1), pp. 6-12.

[Spink 2002] A. Spink, O.H. Cenk, "Characteristics of question format web queries: an exploratory study", Information Processing and Management, 2002, 38, pp. 453-471.

[Topi 2005] H. Topi, W. Lucas, "Mix and match: combining terms and operators for successful Web searches", Information Processing and Management, 2005, 41, pp. 801-817.

[Warren 2001] P. Warren, "Why they still cannot use their library catalogues", Informing Science Conference, 2001,
http://proceedings.informingscience.org/IS2001Proceedings/pdf/WarrenEBKWhy.pdf