Searching for information in an online public access catalogue (OPAC): the impacts of information search expertise on the use of Boolean operators

Jérôme Dinet, Monik Favart & Jean-Michel Passerault
Laboratoire Langage et Cognition (LaCo, CNRS FRE 2725), Université de Poitiers, Poitiers Cedex, France

Abstract
Boolean systems still constitute most of the installed base of online public access catalogues (OPACs) in the French universities even if many studies have shown that Boolean operators are not frequently used by ‘non-librarian’ users (by contrast with professional librarians). The first study examined the use of Boolean operators by French university students; In the second study, elaborated to evaluate the impact of information search expertise on this use, Boolean operators are explicitly presented and participants were explicitly invited to use them. We assumed that university students would not frequently use the operators in searching, and that even if they were explicitly invited to make use of them. Results obtained with the first study based on transaction logs analyses confirmed that French university students did not frequently use Boolean operators. The impact of information search expertise, analysed in the second study, compared three levels of expertise: Novice (university students), intermediate (future professional librarians), and expert (professional librarians). Results showed that, even if the three groups were invited to use Boolean operators, this use increased significantly with the level of information search expertise. University students, if they manage procedural functions of connectives in natural language, do not always manage the whole set of procedural functions carried by such connectives when used in the documentary language. So, the relevance of presenting explicit Boolean operators in the OPACs when users are ‘non-librarians’ is questioned.

Keywords
Boolean operators, cognitive approach, information search expertise, non-librarian users, OPAC

Introduction
Among the various information retrieval systems that can be used in a digital library, On-line Public Access Catalogues (OPACs) are the most widespread in traditional and university libraries. OPACs are mainly concerned with searches for bibliographic records, searching at the level of items rather than full-text content. Since their inception in the early 1960s, OPACs have used Boolean logic to retrieve stored information. Boolean systems still constitute most of the installed base of OPACs in the French universities, even if a lot of studies have shown that Boolean operators are not frequently used by ‘non-librarians’ (by contrast with professional librarians) such as high-school students (e.g., Marchionini et al. 1991, 1993; Sutcliffe et al. 2000). But, in these prior studies, the
‘non-librarians’ were not explicitly invited to use the Boolean operators. So, one of the main questions addressed in this paper is the following: Do university students, novices in information search, use Boolean operators in their queries when they are explicitly encouraged to use them?

Two studies were conducted: The first one investigated the use of Boolean operators by French university students and the second one examined the impact of information search expertise on this use. We assume that even the use of Boolean operators is explicitly required in OPACs and if the users are explicitly invited to use them, ‘non-librarians’ users who have not received a formal training about instructions carried by these operators will not include them in their queries. In other words, the relevance of presenting explicit Boolean operators in OPACs is undermined if users have not previously received formal training in the use of these operators.

Related studies

Large and Behesthi (1997) identified and analysed studies published since 1990. An extensive corpus of writing on OPACs has been accumulated over the past 35 years, even pre-dating their actual implementation in libraries. According to Large and Behesthi (1997), one of the most important difficulties for the non-librarian users was that library OPAC users themselves are not a uniform breed, but come in many guises. If some students may require a considerable expertise in information retrieval tasks in general (and in particular to the use of the individual OPAC system at hand), the majority of students may be novices in both aspects. As Beaulieu and Borgman (1996, p. 492) stated, OPACs ‘are utilised by an extremely heterogeneous user population, probably the broadest of any type of information retrieval system’. The OPACs offer university students the opportunity to consult more information sources than ever before. For the most part, students approach the search for information on the basis of a known item (author or title search) or a topic (Rodman 2000).

According to the Mann’s Principle of Least Effort (Mann 1993), most researchers (even ‘serious’ students) tend to choose easily available information sources, even when they are objectively of low quality. Then ‘non-librarian users’ will tend to be satisfied with what can be found easily, in preference to pursuing high-quality sources whose use requires a greater effort. In other words, end-users such as students want their information search to be quick, easy, usable, and limited in the number of items retrieved. According to experts and practicing librarians, Boolean operators must be used to combine keywords when searching electronic databases, because using these operators might lead to a more focused, quick and easy search.

A lot of empirical studies and cases analyses showed that ‘non-librarians users’ do not frequently use Boolean operators. Several studies have been conducted to collect global data on information search behaviour in a variety of digital libraries (e.g., Connell 1995; Jansen et al. 2000; Spink et al. 2001). Some of these researches have specifically investigated the use of Boolean operators: They attest that ‘non-librarian users’ such as children, high-school students and novices adults have difficulties using Boolean operators, and therefore formulating queries on the basis of these operators (e.g., Hildreth 1982, 1983; Borgman 1986; Siegler 1986; Lietzsch & Marchionini 1988; Case et al. 1989; Marchionini et al. 1991; Rousseau et al. 1998; Sit 1998; Lazonder et al. 2000; Sutcliffe et al. 2000).

Similar results were obtained when the information search task was not performed with an OPAC, but with a search engine in the World Wide Web (Silverstein et al. 1999; Jansen et al. 2000; Spink et al., 2001, 2004; Spink & Ozmutlu 2002; Ozmutlu et al. in press). These studies also showed that most Web users enter few queries including few search terms, conduct little query reformulation and have difficulty using Boolean operators. Many Web end-users recruited in these studies find Boolean logic difficult to master as a form of query construction and many failed to create.

Some studies pointed out that the use of Boolean operators is influenced by the information search expertise (Marchionini et al. 1991; Sutcliffe et al. 2000; Lazonder et al. 2000): Experts in information searching produced more queries and used more specific tools (i.e., Boolean operators and truncation) than novices, even if opposite results were found in one experiment (Marchionini et al. 1993). But, even if these studies provide an insight into the general public practices, some theoretical and methodological limitations prevent us from making any generalisations based on their results. For instance, only a few participants were generally involved in either the case analyses or the empirical studies interested by the use
of Boolean operators in OPACs. Moreover, differences between domain and information search expertise were not clear-cut.

Documentary query vs. natural language (NL)

Why are Boolean operators rarely used by novices (i.e., ‘non-librarian users’) such as university students? According to information search experts and practicing librarians, Boolean logic is extremely easy to use. When used to construct a search expression, it can be very useful for specifying exactly which information is needed by the user:

- In the Boolean language, a search for ‘social AND science’ will locate all the records including both the word ‘social’ and the word ‘science’. In other words, a search for ‘social AND science’ will locate items about the ‘social science’, and the results will exclude those documents containing only ‘social’ or only ‘science’ information. If too many records are retrieved on a topic, a search term can be added with the AND connective. In other words, AND only finds documents including the set of the specified words;
- A search for ‘social OR science’ will locate all the records containing either the word ‘social’ or the word ‘science’, but not necessarily both. If too few records are retrieved on a subject, another search term can be added with this OR connective. OR finds documents including at least one of the specified words;
- A search for ‘science NOT social’ will locate all the records only including the word ‘science’ and will exclude ‘social science’. If too many records are retrieved on an unrelated topic, a term can be suppressed with the NOT connective. NOT finds documents excluding the word mentioned after NOT. In other words, according to the Boolean language, AND implies an exclusion, OR implies a union and NOT a restriction.

In NL, the same connectives, especially AND and OR, imply different ‘rules’. In the NL, AND implies a union. If somebody in a restaurant asks for chocolate AND coffee, s/he hopes to get both. In the same NL, OR implies exclusion: If somebody wants to drive a car, s/he must choose between drink OR drive. According to the context (documentary query vs. NL), AND implies an exclusion vs. a union whereas OR implies a union vs. an exclusion as Table 1 shows.

So the ‘rules’ carried by connectives depend on the context. Caron (1987) and Favart and Passerault (1999) claim that connectives such as AND and OR imply different procedural functions, according to the context, but that users do not always manage the whole set of these procedural functions. In other words, in the case of AND and OR operators, it seems that users such as students do not know which ‘rules’ to apply when searching for information in an electronic database, whereas the same users can manage the ‘rules’ related to these operators in a natural situation (i.e., NL). This suggestion is relevant with the statements of Connell (1995), who mentioned that searching in digital libraries requires a considerable knowledge about syntactic and semantic structures, Boolean operators, and controlled vocabulary.

Do French university students use Boolean operators in their queries when they are invited to use them? When users know they are participants of a research investigation, the drawbacks of methods such as interviews, questionnaires, observation, and thinking aloud protocols centre upon the reliability of the users’ responses or even the normality of the searching behaviour (Large & Beheshti 1997). Users’ keyboard actions as well as the system’s responses captured at the OPAC as a log of each user’s transactions can avoid such drawbacks. A great deal of data can be collected, involving no effort on the researcher’s part, and this data collection being transparency for the users who are not influenced by the presence of researchers hovering around the OPAC (e.g. Pharo & Järvelin 2004). So a first study was conducted to investigate the use of Boolean operators by French university students, by analysing the transaction logs.

Study 1

An investigation was carried out to analyse the transaction logs of a set of 2374 queries produced by 836 university students of an OPAC from September

Table 1: ‘Rules’ of AND and OR according to the language

<table>
<thead>
<tr>
<th>Operators</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Natural</td>
</tr>
<tr>
<td>AND</td>
<td>Union</td>
</tr>
<tr>
<td>OR</td>
<td>Exclusion</td>
</tr>
</tbody>
</table>

2001 to December 2001. Figure 1 shows the search screen of the OPAC used by students. This search screen displayed eight pre-established dialog boxes including Boolean operators (AND, OR, and NOT) explicitly fixed inside.

Two types of analyses were conducted: We first examined the number of terms used per query; second, the percentages of queries including at least one AND, one OR, and one NOT were computed.

As shown in Table 2, the queries were very short: One query contained on average 2.18 terms and 56% of the queries included only one unique ‘search term’: This unique search term could be one term effectively (i.e., one word) or could be the result of the combination of several terms located in the first dialog box. So, even if some users produced several search terms in the same dialog box (generally the first one) these queries were computed as queries including one ‘search term’.

As indicated in Table 3, French university students did not frequently use Boolean operators: Only 28.17% of queries included at least one Boolean operator. Moreover, the operator mainly used was AND and less than 5% of the queries contained OR or NOT.

One limit of this study was that a given search could not be allocated to a specific user. Therefore it was impossible to investigate the relationship between users’ characteristics (e.g., level of expertise) and searching behaviour or to identify the user’s ultimate objective for the search. In other words, if transaction logs provide highly detailed information about how users interact with an OPAC, it cannot give information about the impact of individual characteristics, such as level of information search expertise. For such reasons, transaction logs typically must be used in conjunction with other methods such as experiments.

So, we performed an experimental study including a large sample of participants, to validate the following hypothesis: Even if Boolean operators are explicitly presented and novice searchers (such as university students) are explicitly invited to use them, they will not. If these operators are not used, it is because novice searchers do not know which ‘rules’ to apply when searching for information with an OPAC.

In other words, we sought to answer the two following research questions:

1 Do university students, novices in information, search use Boolean operators in their queries when they are explicitly invited to use them?

2 Does the level in information search expertise affect the use of Boolean operators?

**Study 2**

Participants were provided several search topics for which they were asked to find relevant records using an assigned screen with Boolean operators (see Fig. 1). The impact of information search expertise was analysed by comparing three levels of information search expertise: Novice (university students), intermediate (future professional librarians) and expert (professional librarians). Three indicators were analysed for each expertise level: (1) the number of queries produced, (2) the percentage of these queries including at least one Boolean operator and (3) the nature of the used operators (AND, OR and NOT).
Method

Participants
One hundred and seven participants took part in this study. Three groups were constituted according to their level of information search expertise. The expert group included twenty-two professional librarians (mean age = 35.2 yr) who had been working for at least ten years in libraries in several secondary schools. The intermediate group included eighteen students (mean age = 23.9 yr) preparing the librarians’ exam. They had received one year of formal training in electronic information searching. The novice group included sixty-seven university students who were all novices in information search: Thirty-eight (mean age = 25.9 years) were fourth-year students from the department of literature and 29 (mean age = 24.5 years) were fourth-year students from the department of medieval history. The impact of knowledge expertise was not investigated, but it was controlled by using six search topics related to two main subjects (modern literature vs. medieval history) and to recruit university students in two academic courses (history vs. literature).

Task and materials
Participants were first told they had to perform queries to find bibliographic records about six topics related to two main subjects. Three topics (police novels, French surrealism and the Annals school) were related to modern literature. Three other topics (knights’ armour, gothic art and crusades during the Middle Ages) were related to medieval history.

As in study 1, each participant was asked to perform her/his queries using an electronic screen copy. This electronic screen copy displayed eight pre-established dialog boxes including Boolean operators (AND, OR, and NOT) explicitly fixed inside (see Fig. 1). So, each participant was asked to compose queries on the screen copy, writing search terms in each dialog box, and s/he was explicitly invited to use Boolean operators to ‘connect’ the search terms produced.

Procedure
Participants were given the six search tasks topics for which they were asked to find relevant records using an assigned screen (with the following Boolean operators: AND, OR, and NOT). One of the six topics was presented to each participant (police novels, French surrealism, the Annals school, knights’ armour, gothic art or crusades during the Middle Ages). S/he was asked to perform queries to find bibliographic records about this topic. S/he was explicitly informed that the use of the connectives AND, OR, and NOT was possible and explicitly recommended. The exact instructions were as follows:

Imagine you have to search for bibliographic references about police novels in your traditional [or university] library. You must use an electronic tool to carry out this bibliographic search. The interface of this electronic tool [Fig 1 was presented] displays several pre-established dialog boxes into which your search terms can be entered. The contents of these different dialog boxes can be connected by Boolean operators ‘AND’, ‘OR’, and ‘NOT’. In other words, when you write your queries you are invited to use one, two, or more Boolean operators to connect the search terms you produce.

The participant could ask question to the experimenter. If s/he had no questions and said s/he was ready, the experiment could begin.

Once the participant had performed the first query, s/he was invited to generate a second query on the same topic, using an identical electronic screen copy. After performing this second query, s/he was invited to generate a third one and so on. This meant that each participant could compose as many queries (one, two, three, etc.) as s/he wished on the same topic.

When a participant informed the experimenter s/he had no more idea how to compose a new query about this first topic, a second topic from the same theme (modern literature vs. medieval history) was provided and so on. The procedure applied to the following topics was the same as for the first one. The same procedure was used for the second theme (modern literature vs. medieval history).

The experiment was completed either when the participant had produced queries for all three topics for each theme (modern literature vs. medieval history) or when s/he wanted to stop. Therefore, no time limit was set for any search topic.

Results
For each level of information search expertise (novice, intermediate, and expert), we analysed the number of queries produced, the percentage of these queries including at least one Boolean operator, and the percentage of these queries including AND, OR and NOT.
As Table 4 shows, the number of queries of the participants was approximately identical for the three levels of information search expertise. Data were entered in a 3 (Search expertise: novice, intermediate and expert) × 2 (search topics: literature, history) ANOVA with the first factor as a within-subject factor and the second one as a between-groups factor. They attested that the number of queries produced was not significantly influenced by the level of information search expertise (novice = 1.42; intermediate = 1.48; expert = 1.28; F(1–104) = 2.42, P = .09). We did not observe interaction between the search expertise and the search topic.

Across the search topics, the percentage of queries including at least one Boolean operator increased with the level of information search expertise (novice, 32.09%; intermediate, 58.44%; expert, 82.53%). χ² attested that the experts produced significantly more queries including at least one Boolean operator than the intermediate group ($\chi^2 = 8.92, P = 0.003$). Moreover, the intermediate group produced significantly more queries including at least one Boolean operator than the novices ($\chi^2 = 8.43, P<0.02$).

χ² tests revealed that the percentage of queries including the AND operator was always higher than the percentage of queries including the OR operator (respectively, 77.99% and 10.37%; $\chi^2 = 92.90, P<0.0001$) and the percentage of queries including the NOT operator (11.63%; $\chi^2 = 34.66, P<0.0001$), while the percentage of queries including OR and NOT was not significantly different ($\chi^2 = 0.08, P = 0.77$). No other significant result was observed.

In other words, the nature of the Boolean operators used was not determined by the information search expertise. Although the level of information search expertise did not influence the number of queries performed by participants, it significantly influenced the percentage of queries including at least one Boolean operator, which increased with the level of information search expertise. This experiment also showed that a third of university students (the novice group) did not use Boolean operators even when these operators were explicitly mentioned and when they were invited to use them.

**Discussion**

Two studies were conducted to examine, respectively, the use of Boolean operators and the impact of information search expertise on this use. In both studies, the instructions and/or the material (i.e., the search screen) explicitly focused on the use of Boolean operators. Results obtained in study 1 confirmed results of previous studies: University students do not frequently use Boolean operators, and showed the mainly used operator is AND. In study 2, also convergent with the findings of some prior studies (Marchionini et al. 1991; Lazonder et al. 2000; Sutcliffe et al. 2000), information search expertise significantly affected the number of queries including at least one Boolean operator: The use of Boolean operators increased with the level of search expertise.

First, participants in the intermediate group used Boolean operators more frequently than the participants in the novice group. The main difference between the novice (university students) and intermediate (future professional librarians) groups was learning through formal teaching of the use of Boolean operators. Participants in the intermediate group had therefore become more expert in the use of

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Novice (n = 67)</th>
<th>Intermediate (n = 18)</th>
<th>Expert (n = 22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean number of queries produced</td>
<td>1.42 (0.48)</td>
<td>1.48 (0.25)</td>
<td>1.28 (0.41)</td>
</tr>
<tr>
<td>Percentage including at least one Boolean operator</td>
<td>32.09</td>
<td>58.44</td>
<td>82.53</td>
</tr>
<tr>
<td>Percentage of AND</td>
<td>79.25</td>
<td>78.77</td>
<td>76.2</td>
</tr>
<tr>
<td>Percentage of OR</td>
<td>16.55</td>
<td>17.65</td>
<td>12.82</td>
</tr>
<tr>
<td>Percentage of NOT</td>
<td>4.19</td>
<td>10.32</td>
<td>10.92</td>
</tr>
</tbody>
</table>

Boolean operators than the novices through formal academic training in electronic searching. In other words, results are consistent with the following assertion: French university students, who are novices in information search, did not frequently use Boolean operators even when they were explicitly invited to use them, because they did not manage the specific ‘rules’ carried by these operators in the documentary context. So, searching in OPAC requires a considerable knowledge about syntactic and semantic rules, these rules being unknown to the ‘non-librarian’ users such as university students.

Second, experts (professional librarians) in information searching used Boolean operators more frequently than the intermediate group. The main difference between the intermediate and expert groups was experience, in the sense of practice. The expert group included professional librarians who had been working in different libraries for at least 10 years. These librarians had received formal teaching in the use of Boolean operators (like the intermediate group) but had also practiced electronic information searches for at least 10 years. In other words, although formal teaching was necessary to learn how to use Boolean operators, their use was also influenced by experience and practice.

If the results of study 2 confirm the findings of earlier studies by showing that the use of Boolean operators is essentially influenced by the information search expertise, future research should focus on the factors which affect ‘non-librarian users’ performance and seek ways to help these end-users to improve their performance. A first hypothesis could be tested in future investigations: The easier the use of search engines is (i.e., without Boolean operators), the more ‘non-librarian users’ will use Boolean operators. Indeed, as Owei (2000) argued, the focus in database systems needs to shift more in the direction of interfaces adapted to the natural mode of human communication. In this sense, researchers in NL processing and computational linguistics have proposed or developed various approaches to NL interfaces for human–machine interaction, without Boolean operators. The underlying idea is that with NL interfaces, there is no need for users to learn a new syntax and new ‘rules’. With NL interfaces, users are not subjected to rigid syntactic rules, and common deviations from formally correct grammar could be tolerated as long as the intent of the query is clear. However, one of the main problems is largely due to the inability of these NL interfaces to deal with subtleties in the human use of NL. And even if some researchers have begun to try and to solve some types of problems (Lewis & Jones 1996; Wiebe et al. 1996; Owei 2000), computers are not really able of catering for linguistic subtleties. Studies comparing the performance of users with NL systems or with other systems have yielded mixed results. While some studies have reported the superiority of NL interfaces (Vassiliou et al. 1983; Suh & Jenkins 1992; Chan et al. 1999), others have shown that NL users performed more weakly (Small & Weldon 1983; Jarke et al. 1985). Others have found no significant differences (Turner et al. 1984).

A second hypothesis could guide future investigations: Results of study 2 indicate that the relevance of presenting explicit Boolean operators can be undermined if users have not previously received formal training in the use of these operators. But the most common reason that end-users gave for not using the OPACs is that they had not time to learn how to use it (Matthews et al. 1983; Borgman 1999). Thus, according to Borgman (1999), an implementation problem exists: We have several complex systems to be learned, and end-users without time and motivation to put some effort into learning to use them. We therefore need an efficient method of teaching novice end-users how to use these technologies.

Several methodological limitations prevent us from making generalization based on the data obtained in the second study. First, we did not evaluate frequency of library usage. Measures of frequency of catalogue use and of knowledge of the catalogue tool would have been helpful. They should be included in future research even if we do not know the exact purpose of the library visits (they could be to retrieve information or merely to find a quiet refuge for study) and even if a recent study has found that experience in using a library search tool does not affect the search performance (Fang & Salvedy 2000). Second, we didn’t focus on the effectiveness of the queries: In a further experiment, we should examine this effectiveness.

Acknowledgement

The authors would like to thank Thierry Olive for his careful reading and helpful comments.
References


