Carmen Caro Castro and Críspulo Travieso Rodríguez
Department of Library and Documentation Science. University of Salamanca, Spain

Ariadne´s Thread: Knowledge Structures for Browsing in OPAC´s

Abstract: Subject searching is the most common but also the most conflictive searching for end user. The aim of this paper is to check how users expressions matche subject headings and to prove if knowledge structure used in online catalogs enhances searching effectiveness. A bibliographic revision about difficulties in subject access and proposed methods to improve it is also presented. For the empirical analysis, transaction logs from two university libraries, online catalogs (CISNE and FAMA) were collected. Results show that more than a quarter of user queries are effective due to an alphabetical subject index approach and browsing through hypertextual links.

1. Introduction

Since the 1980’s, online public access catalogs (OPAC’s) have become usual way to access bibliographic information. During the last two decades the technological development has helped to extend their use, making feasible the access for a whole of users that is getting more and more extensive and heterogeneous, and also to incorporate information resources in electronic formats and to interconnect systems. However, technology seems to have developed faster than our knowledge about the tasks where it has been applied and than the evolution of our capacities for adapting to it. The conceptual model of OPAC has been hardly modified recently, and for interacting with them, users still need to combine the same skills and basic knowledge than at the beginning of its introduction (Borgman, 1986, 2000): a) conceptual knowledge to translate the information need into an appropriate query because of a well-designed mental model of the system, b) semantic and syntactic knowledge to be able to implement that query (access fields, searching type, Boolean logic, etc,) and c) basic technical skills in computing.

At present many users have the essential technical skills to make use, with more or less expertise, of a computer. This number is substantially reduced when it is referred to the conceptual, semantic and syntactic knowledge that is necessary to achieve a moderately satisfactory search. An added difficulty arises in subject searching, as users should concrete their unknown information needs in terms that the information retrieval system can understand. Many researches have focused on unskilled searchers’ difficulties to enter an effective query. The mental models influence, users assumption about characteristics, structure, contents and operation of the system they interact with have been analysed (Dillon, 2000; Dimitroff, 2000). Another issue that implies difficulties is vocabulary: how to find the right terms to implement a query and to modify it as the case may be. Terminology and expressions characteristics used in searching (Bates, 1993), the match between user terms and the subject headings from the catalog (Carlyle, 1989; Drabenstot, 1996; Drabenstot & Vizine-Goetz, 1994), the incidence of spelling errors (Drabenstot and Weller, 1996; Ferl and Millsap, 1996; Walker and Jones, 1987), users problems
to find alternative terms and to reformulate the query and the movements sequence of the modification (Ferl and Millsap, 1996; Spink and Saracevic, 1997) have been examined.

Results of these researches seem to bring in question the effectiveness of analytic strategies for those end-user queries where the information needs have not been clearly defined. In these cases, improving effectiveness involves creating alternative models in order to reduce the cognitive effort that is required to formulate a search strategy that the information system understands. They represent methods that facilitate some kind of browsing and that imply a dynamic and exploratory searching (Bates, 1989, Marchionini, 1997).

In this context, knowledge structures used traditionally in information systems (classifications, subject headings and thesaurus) have not been considered nowadays as an obstacle, as it has been asserted before by the proponents of automated systems with natural language indexing; instead of this, they have come to be considered as effective tools, combined with advanced techniques of information retrieval (Chan, 2000). Reasons for this new vision are that these tools enhance the conceptual knowledge of users about the information system, its contents and how the information space has been organized, facilitating alternative terms to extend or modify the query (Efthimiadis, 1996). Besides, these established tools are useful for mapping the user's terms to a controlled vocabulary (Bates, 1998, Buckland et al., 1999).

In Spain, empirical studies about end-users subject searching in OPAC's have not been developed. Nevertheless, there are exploratory studies, like the one who analysed the relevance of results from 25 subject searches in CSIC libraries network (González Sereno & Soria González, 1996); another study dealt with a revision of new targets for improving the automated subject access with OPAC's (Seguí i Palou and Vall, 1999). Besides, an implicit problem derives from the non-existence of a Spanish list of subject headings equivalent to Library of Congress Subject Headings or Vedettes-matière of University of Laval. Owing to this circumstance, libraries have fallen back on different methods to fit the subject headings to their own needs, generating both problems of consistency of the representation and the lack of an effective and consistent syndetic structure for providing better subject access in information networks (Gil Leiva, 2001, Jiménez Rodriguez, 1998).

This research work analyses queries in a subject field entered by OPAC end users from a university library, in order to check the match with the terminology used in the indexes of the online catalogs from two libraries which work with different subject headings lists: the Universidad Complutense de Madrid (CISNE) and the Universidad de Sevilla (FAMA). Specifically, this work is aimed at the following objectives:

- to know about the characteristics of searching statements: typology, errors, structure.
- to check the degree of matching between user terminology and the subject headings from both online catalogs.
- to analyse the relationship between characteristics of user queries (MARC fields, number of words, syntactic structure) and the match with both controlled vocabularies.
to examine syndetic structure of the controlled vocabulary, alphabetical subject indexes and hypertextual links of bibliographic records in order to measure their usefulness for retrieving records.

2. Methodology

The subject search statements were extracted using transaction logs dated March to May, 2001 from the Universidad Complutense Library's online catalog (CISNE). The number of queries was 34,908; from this amount were excluded: a) duplicated queries originated at the same day, at the same terminal and with a two minutes interval; b) non legitimate queries – n, r; njhjh - ; c) queries in foreign languages – cognitive neuroscience, decision making -. From the remaining number a random sample was collected for a $\alpha = 0.05$ and 95% precision; overall, the obtained sample consisted of 385 subject search statements, which are the ones analysed in this research work.

In the first place, we examined the characteristics of search statements, such as typology according to MARC fields, number of words per statement, and syntactic structure. For establishing structure categories it all the possible ones were counted, and those categories whose frequency was lower than 1% were included in a unique category called Others.

The next step was to check if there were spelling or typographical errors in the subject queries. Even though it have been established different categories for spelling errors (Drabensttot and Weller, 1996), the types considered in this research work were substitution, transposition, insertion and omission of characters or blanks. These errors were corrected before searching so as to determine what it would be the system answer if there were no errors. The same procedure was applied in the cases of singular/plural variations and correct spelling variations.

At the time of searching, queries were entered in the OPAC Web catalog from the cited institutions, which is managed by INNOPAC Millenium. This system allows to search on rotated subject headings indexes and by keyword matches. In order to search by keyword in subject headings matches was necessary to use REBIUN Web OPAC (Red de Bibliotecas Universitarias), that provides an independent access to each integrated library's collection.

The variables that measure the matching between user expressions and subject headings of the information system are a representation of degrees of similarity, related to terminology and syntax. Although many researchers have undertaken this kind of analysis, the variables that have been used are so heterogeneous that it is complicated to make comparisons fairly. At this point, it was decided to establish categories similar to the ones used by Carlyle (1989), Drabensttot and Vizine-Goetz (1994), Drabensttot (1996) and Drabensttot and Weller (1996); these variables were the closest to our objectives and the most effectively comparable. Despite this, punctuation nor upper/lower case letter variations were considered in this study.

Each user query was searched to see if it fell into the first match category; only when this matching failed it was checked in the next category. Consequently, it is very important to bear in mind the next ordered sequence of categories used in this paper:
1. Subject heading matches.- First of all, to check how similar user vocabulary and syntax are to both controlled vocabularies, two categories of matching were used:
   - Exact match.- Subject heading matches user expression exactly.
     In this section there are two variants:
     ✓ Word order variation.- Subject heading contains the same words as user expression, but not in the same order.
     ✓ Exact substantive word match.- Subject heading contains the same substantive words which are in the user expression, although word order may differ.
   - Partial match.- Subject headings consists of the same substantive words which are in the user expression, although it includes one or more additional terms.

   We corroborated the relationship between variables such as type of heading, structure and number of words with the variable exact matching. To verify this relationship we used the chi-squared test, significant for \( p<.005 \).

2. Alphabetical approach.- In case queries contained words not present in subject headings, the alphabetical index was examined. INNOPAC Millenium system sets user query at the corresponding place in the alphabetical subject index. This option may be useful for user to find an alternative subject heading to continue the search. Based on the first alphabetical index display offered by this system - 12 headings -, we distinguished three possibilities: subject heading may be equivalent, more specific or more generic than user expression.

3. Keyword matches.- At last, those user expressions which did not fall into the preceding categories were searched as keyword in record. We checked the relevance of the first ten retrieved records and verified whether they contained alternative subject headings to lead the user to find relevant records using hypertextual links.

3. Results

   The percentage of topical subject searches (MARC tag 650) was higher (86%) than the percentage of the same category obtained by Drabensttot and Vizine-Goetz, 1996 (69.5%); therefore, the rest of fields - 600 (5.7%), 651 (3.9%), 610 (3.12%), 630 (1.3%) – happened in lower proportion than in the cited study. Regarding the number of words by statement, the mean was 2.24 words (ranked from 1 to 8); this figure is slightly higher than the one presented by Drabensttot and Vizine-Goetz (1996) -1.8 words- and by Drabensttot (1996) – 1.6 words- . In relation to number of words, 37.4% of the queries just consisted of a single word, 30.65% consisted of two, 15.06% of three and 17% of more than three. That distribution is directly related to the syntactic structure, as the most common structures are one term (common noun) – 30.13% , noun and adjective – 19.74% , and proper noun – 11.43% , which is usually a surname, or a surname and first name.
As table 1 shows, the syntactic structure of the statements coincides for the most part with natural language and with subject headings without subdivisions: common noun, common noun and adjective, proper noun, noun and prepositional complement. However, it is observed that some users entered statements not in natural order; this may imply a certain adjustment to subject heading syntax or use of headings proposed by the system - marruecos cultura, juegos educacion fisica, literatura españa -. Due to the differences of categories, comparison with Bates (1989) results is not feasible.

The percentage of searches with spelling or typographical errors (Table 2) was 6.5%; 52% of them were omission errors. This result is higher than the percentage reported in Blecic et al., 1999 (3.18%), Jones, 1996 (5%) and Drabenstot and Weller, 1996 (5.9%). Singular/plural variations between queries and subject headings represented 3.5% in both universities.

In the analysis of exact matching between user expressions and subject indexes, we found differences for each catalog of the two centres; there are two possible reasons. On one side, users may have adapted their queries depending on the catalog in use; on the other side, user terms may have been derived from used terms in other searches, since transaction logs proceed from Universidad Complutense, where the percentage of exact matching is higher. At Universidad de Sevilla, exact matching – 31.95% - is not very affected by normalization and it is not affected at all by See references. As opposed to this, in Universidad Complutense catalog exact matching increases 3.54% due to normalization and 3.64% due to cross references, obtaining a total percentage of 45.2% for exact matching. In both cases, these results are lower than those reported by Carlyle, 1989 (60%) and Drabenstot and Weller, 1996 (48.5%), but higher than Drabenstot and Vizine-Goetz, 1994 (33%).

<table>
<thead>
<tr>
<th>Correct spelling variation</th>
<th>2</th>
<th>0,52%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling errors</td>
<td>2</td>
<td>0,52%</td>
</tr>
</tbody>
</table>

**Typographical errors**

| Substitution | 5 | 1,30% |
| Transposition | 2 | 0,52% |
| Insertion     | 5 | 1,30% |
| Omission      | 13| 3,38% |

**Total of searches with typographical errors** 25 6,5%

Table 2. Spelling and typographical errors.

More than 90% of exact matching queries were queries for topical subjects in both catalogs; although this kind of query is the most common, the difference from
the rest of queries was significant (p<.005). However, number of words or syntactic structure are variables with more incidence (p<.001); A fact which speaks for itself is that, in both catalogs, more than 50% of exact matching queries was one word statement, and 30% was two words statements; this relationship was also verified by Markey (1984). In addition, more than 50% of exact matching queries consisted of common noun, and 20% consisted of common noun and adjective.

Table 3 also lists the number and percentage of partial matching: 20.78% in Universidad Complutense and 17.14% in University of Sevilla. This variable comprises those user queries contained in a subject heading that included one or more additional terms (for instance, a geographic subdivision). In these cases, the possibility to retrieve relevant records is reduced; because of it, systems must offer a list of such headings so that it stimulates users to narrow or focus their topics of interest. The obtained figures are higher than Carlyle, 1989 (10%), and also slightly higher than Drabenstott & Vizine-Goetz, 1994 (16.2%) and Drabenstott & Weller, 1996 (14.7%). Exact and partial matches bring users back a list of potentially relevant terms or records, whenever the system allows us to remove non-substantive words and to search by keyword in heading matches or in rotated indexes. But, it is necessary to know what happens to those queries that do not fall into these matching categories; in other words, it is the right time for a subject headings list to prove its effectiveness in searching. And, as table 4 summarizes, the percentage of alternative headings that are provided by alphabetical subject index is higher in both universities than the same percentage provided by nonnalization. Besides, in Universidad de Sevilla, where exact matching was lower, this figure means a considerable number of queries.

Regarding to searches by keyword matches, 55.11% and 75.58% of these searches provided relevant records in Universidad de Sevilla and Universidad Complutense, respectively. From those retrieved records, over a half in Universidad de Sevilla (56) and about 2 of every 3 in Universidad Complutense (73) contained at least one alternative heading useful to continue the search or to browse the catalog with hypertextual links. This figures represent 15% (Sevilla) and 19% (Complutense) of the total of queries statements. Summing up, if we add these percentages to the results obtained in the preceding category (alphabetical approach), we find that possibilities to find relevant records with the help of browsing through system tools (links between records indexed with the same headings) and alphabetical subject heading index account for more than 25% in both university online catalogs.

<table>
<thead>
<tr>
<th></th>
<th>Univ. of Sevilla</th>
<th>Univ. Complutense</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>%</td>
</tr>
<tr>
<td><strong>Exact matching</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct</td>
<td>115</td>
<td>29,87</td>
</tr>
<tr>
<td>See reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word order variation</td>
<td>2</td>
<td>0,52</td>
</tr>
<tr>
<td>Exact substantive word match</td>
<td>6</td>
<td>1,56</td>
</tr>
<tr>
<td><strong>Total of exact matching</strong></td>
<td>123</td>
<td>31,95</td>
</tr>
<tr>
<td><strong>Partial matching</strong></td>
<td>80</td>
<td>20,78</td>
</tr>
</tbody>
</table>

Table 3. Exact and partial matching.
4. Conclusion

In the matter of characteristics of queries, we found an extensive range of structures, but the most common was topical subject and with only a few words. Results analysed above show that queries with more probability to be exact matches are those with simple structure and few words. The percentage of queries with typographical or spelling errors or singular/plural variations accounted for 11%; to solve this problem, a system of detection of misspelled words (Drabensttot and Weller, 1996; Walker and Jones, 1987) could be implemented.

To improve the match between user statements and controlled vocabularies it would be expedient to incorporate into online catalogs new techniques of information retrieval; these should be combined and promoted in a way like the search trees proposed by Drabensttot (1996). It has been observed above that subject heading lists are a very valuable tool for enhancing access to information: an improved syndetic structure and crossed references provides alternative terms and increases the proportion of exact matches. Besides, linking records which are indexed with the same subject headings brings users the possibility of effective browsing.

Alphabetic subject heading indexes may become a important instrument for guiding user through the library collection, focusing or narrowing users queries, as Ariadne’s thread guided Theseus to find the way out of the maze. That is the essential aim of their knowledge structure, which can and should be developed by technology (hypertext had opened new ways for retrieval) and by consolidated indexation instruments (alphabetical indexes and references).

Notes
1.- To calculate the sample size it was used GRAMNO 4.0, and to select queries randomly and the rest of statistical process it was used SPSS 10.0.

References.


