Analysis of Subject Searching in the Tenttu Books Database

Irma Pasanen-Tuomainen

Helsinki University of Technology
BACKGROUND OF THE PROJECT

This study was the pilot study for an Internordic project to monitor the use of online catalogues in the Nordic technological university libraries to be started in the Fall 1991. Most of the Nordic online OPACs have been in operation for more than a decade but not much information is available on how the users actually carry out their searches in these systems. On the whole, subject searching in OPACs is generally little examined. In this pilot study, the main emphasis was on the use of classification in subject searching: how the Universal Decimal Classification (UDC) is used and the extent of its use. The aim of the Internordic project is not only to gather information about user interaction with the OPACs and to focus on subject searching as well as to consider the improvements to information retrieval in the catalogues. The method chosen to carry out this project was the transaction log method, a rather inexpensive way to gather information.

There have been several surveys concerning the TENTTU Books database in the past ten years. None of them monitored the actual searches as they were based on interviews and questionnaires. The problem with questionnaires is that the answers can be to a degree misleading [1], and therefore the results of such surveys must be carefully scrutinized.

The Helsinki University of Technology Library is now installing an integrated online system in addition to the TENTTU system. For the Internordic study, searches made in the new Virginia Tech Library System (VTLS) will be monitored as well to allow comparisons between the two systems.

THE AIM OF THE PILOT STUDY AND THE METHOD USED

The TENTTU Books Database

The TENTTU Books database is the online union catalog of the Helsinki University of Technology Library. It has some 110,000 references to books, reports, conference proceedings, dissertations and parts of series acquired by the main library and the 26 satellite libraries within the university since 1980. It is maintained on a MicroVAX computer along with the other databases of the TENTTU system at the Computing Center of the University. The retrieval system is an application of the Swedish software TRIP, originally developed at the Royal Institute of Technology Library. The search language is CCL (Common Command Language) and the system operates in a command-mode only.

TENTTU Books is a multilingual database, i.e. all the references are put in the system in their original language. The classification used in the system to describe the contents of an individual publication is based on the Finnish abridged version of the Universal Decimal Classification with special emphasis on divisions 5 (Mathematics and natural sciences) and 6 (Applied sciences, medicine and technology). Each publication is indexed generally with two to four classification codes (Figure 1). Some characteristics of the UDC are not used in the TENTTU Books database, characteristics that were not considered to promote retrieval. The UDC tables and alphabetical index have not been online since the change of the software of the TENTTU system in the mid 1980's.

TENTTU Books is not only an online catalog, but also a true information retrieval system. All fields are searched with the search command FIND,
only the use of the field prefix limits the search to a certain field. Since no index terms are used to describe the contents of a publication, the search for a specific subject must be conducted with the use of the classification or with free text searching of titles and some other parts of the reference.

The Monitored Users

There were three different groups of users monitored (all that is known about an individual user is the user category):

First, the outside clients who are charged according to the connect time in the system. The group consists mainly of staff in other libraries, both academic and public as well as information services of industrial enterprises. It can be assumed that this group has some knowledge of classification in library systems. Whether they have access to printed UDC tables is unknown. There are free TENTTU seminars every first Monday of the month to train these users.

Second, the users of the computer terminal next to the reference librarian’s desk at the library. This group contains mainly students of the Helsinki University of Technology. A printer is available as well as assistance by the information specialist on duty. Also the TENTTU system manuals and the Finnish abridged version of the UDC are next to the terminal. Most of the students have taken part in the courses on information retrieval. These courses are taught by the information specialists in the library.

The third group monitored was the staff of the information service department of the library (excluding those working with classification). This group is expected to know not only the UDC as a method to describe the contents but also how to use it as a tool in subject searching. This group is also expected to know the structure of the TENTTU Books database.

The Method of Analysis

The method used to monitor the searches was the transaction log method. All search sessions in the TENTTU Books database during March 7 to 28th 1991 by the three different user groups were recorded. Search commands were categorized to find out the extent of different types of subject searches. A closer look was taken in those sessions where the UDC classification code was used. The log files were treated with a spreadsheet program of a microcomputer to delete the irrelevant information from the logs and to calculate the results. The categories listed in Table 1 consisted of the searchable fields of the database and an additional free text category. In indistinct cases the context and structure of the whole search sentence helped to determine the category. The use of Boolean operations in a search sentence allowed more than one category to apply in one sentence.

SUBJECT SEARCHING IN THE TENTTU BOOKS DATABASE

Conducting a Subject Search in the Database

The classification code is used to describe the contents of an individual publication in the database and therefore the use of the code in subject searching is encouraged to retrieve the maximum number of references. The bare use of free text terms in subject searching makes it hard to overcome the language barrier in a multilingual database. However, it is also possible to combine classification codes with either free text terms or specific fields.

In the field “Classification Code” (CC) in the database, the UDC codes are displayed and different codes are separated with a colon (:). Both main divisions and auxiliaries are put into inverted files and they can be searched
separately. Due to the structure of the system, neither period (.) nor colon (:) is a searchable sign with UDC-codes. Instead, the sign understrike (_) must be used to separate divisions and Boolean logic is used to combine codes. Truncation is also recommended when searching with classification.

The command FIND CC=629.114$ and a485 and industry would therefore retrieve references to documents about the motor vehicle industry in Sweden but probably in English language only as the term "industry" must be found in the reference.

Results of the Analysis

The Extent of the Use of the UDC

During the monitoring period there were 655 search sessions monitored. These sessions included 2620 executed search commands in the TENTTU Books database, searches that started with the command FIND or F. One third of these searches included a Boolean operator to combine search elements thus adding the total number of given individual search elements to approximately 3500.

The number of searches in different categories is shown in Table 1. As a search command could include elements of different categories the total number of all searches does not equal the number given above. The most popular way to search was using free text terms, 70.8% of all the searches were free text searches. Of these same 11% were also found in the alphabetical index of the Finnish abridged version of the UDC and therefore they can be interpreted also as keywords. The use of the UDC code varied a great deal according to the user group: the information specialists of the library used it in 13.6% of the search commands they gave, but for both the outside clients and students it seemed to have little significance in subject searching. Since TENTTU books is an online catalog, the possibility to locate a publication by the name of the author was often used by the outside clients and students. The option to use the title for the same purpose was much less used. It can be assumed, however, that a large number of searches by the name of the publication fell in the category of free text search because of the structure of the search sentence. The use of Boolean logic is allowed (and encouraged in the TENTTU manuals) in all phases of the search and a title can be searched either with the field prefix TI = or by combining the words of the title with Boolean operators. Field prefixes LA = (language) and PY = (publishing year) were also used to some extent by all of the user groups, their role in the search is usually to restrict the search to certain languages and years.

The Boolean operators were used in almost one third of the searches. Search terms were truncated in 43% of the searches. Errors occurred in 264 (10%) of the searches and they consisted mainly of misspellings. However, only errors made before entering the database or mistakes made with other commands were included in this survey thus excluding mistakes made before entering the database or mistakes made with other commands.

The searches by the UDC code, 3.7% of the total, were all subject searches. If all free text searches (70.8% of the total) are interpreted respectively almost three quarters of all monitored searches in TENTTU Books are classified as subject searches. Searches for an author or title summed up to 19.2%. 

74
A closer look was taken at those 30 sessions where the UDC code was used as a search element. There was no approach that would apply to all these sessions but some similarities could be found. The sessions began in 22 cases with a search for a term and the UDC code was introduced later, either as a restrictive element to the term or in a new approach within the same session. In 23 of the sessions the code or codes described the broad context of the subject which was then narrowed down either with other aspects such as a language or year or with a term. In the remaining seven sessions the UDC code acted as the element to restrict the search of the term. Misspelling the UDC code occurred in nine sessions but it was corrected by the user in the next search command in all but two cases.

In the case of no hits after a search starting with a term the users in their next approach preferred to search for a term again rather than switch to the UDC code describing the subject. Also, if there were no hits after the search with the UDC, more users switched to terms than remained with the UDC code in their next approach.

Conclusions

The extent of the use of the UDC codes during the monitoring period was smaller than expected. No statistical analysis on how they were used in subject searching can be made since the codes appeared in 30 sessions only.

The information specialists of the Helsinki University of Technology Library used the UDC code far more than either students or outside clients. This was expected as the staff works with the system daily and knows the system well. Also the use of the category "other", including searches with rarely used prefixes such as RN = for report number, is a result from knowing the system.

In a recent survey on Finnish databases [2] TENTTU Books was criticized because the UDC classification code is not presented in the manual nor is it available online. This is probably the most important reason for the little use of the code. The survey also noted the possibility of adding keywords to the system to help retrieval and suggested that in the user training offered

<table>
<thead>
<tr>
<th>Category (field prefix)</th>
<th>Number of FIND commands including the search category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clients</td>
</tr>
<tr>
<td>Free text</td>
<td>1011</td>
</tr>
<tr>
<td>Author (AU=)</td>
<td>156</td>
</tr>
<tr>
<td>Title (TI=)</td>
<td>37</td>
</tr>
<tr>
<td>UDC code (CC=)</td>
<td>29</td>
</tr>
<tr>
<td>Conference title (MT=)</td>
<td>6</td>
</tr>
<tr>
<td>Serial name (SO=)</td>
<td>12</td>
</tr>
<tr>
<td>Language (LA=)</td>
<td>4</td>
</tr>
<tr>
<td>Publishing year (PY=)</td>
<td>26</td>
</tr>
<tr>
<td>ISBN (SS=)</td>
<td>5</td>
</tr>
<tr>
<td>ISSN (SS=)</td>
<td>0</td>
</tr>
<tr>
<td>Publisher (PU=)</td>
<td>1</td>
</tr>
<tr>
<td>Printing location (PT=)</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>1309</td>
</tr>
</tbody>
</table>
to the clients there ought to be more emphasis in subject searching with the UDC code.

It might be unclear to some users that in a free text search the term only retrieves the references written in the language of the term in question. The usage of the most effective tool in subject searching in the multilingual database must thus be constantly encouraged. At least the main divisions of the code should be listed in the manual.

COMPARISON TO OTHER STUDIES CONDUCTED IN FINLAND

There has been a debate in Finland in recent years about the problems in describing the contents of publications in online catalogues. Especially now when the Finnish research libraries are all installing the VTLS library system, the topic "UDC code versus index terms" is a subject to much discussion.

In a study concerning the national databases based on transaction logs it was found that the UDC code was used in about 7% of all searches done using the databases KOTI (the Finnish national bibliography) and KAUKO (the union catalog of foreign publications in Finland).[3] This is somewhat more than the use of the code in the TENTTU Books database. Most of the users of these databases are librarians in public and academic libraries familiar with classifications. In the transaction logs of this study the most popular way to search seemed to be with free text terms as in the TENTTU Books database, too.

A project on describing the contents of publications carried out prior to the transaction log study of the national databases based its analysis on interviews and questionnaires among librarians. During the project the users were also trained.[4] In this survey the use of the UDC code in the searches was found to be far more extensive (19% respective 30% of all searches conducted in databases KOTI and KAUKO were done by using the UDC code) than in the study based on transaction logs.

The results of this pilot study in the TENTTU Books database together with the transaction log study in the national databases support the idea presented in the study concerning the translation of UDC codes into index terms [5] that the use of classification codes in subject searching in library catalogs in Finland is fading and people prefer words to numbers when structuring their queries.

PROJECT EVALUATION

This pilot study fulfilled its task in terms of gathering important information for the forthcoming Internordic study. The first finding here is that there must be far more search sessions monitored than the 655 done in this pilot project to gain a statistically relevant amount of searches with the UDC-code. Secondly, criteria to analyze the free text searches must be set up to determine the share of true subject searches within the free text searches.

When the new VTLS integrated library system is introduced to the users at the Helsinki University of Technology Library the implications of the conclusion that the UDC code is not used in subject searching as much as expected have to be discussed.

There was an experiment done in Finland to translate the Finnish abridged version of the UDC into index terms.[5] Also, there is an ongoing project to update and enlarge the alphabetical index of the Finnish abridged version of the UDC.[6] The multilingual TENTTU Books requires index terms not only in Finnish but also in English, if the clients are to use the database effectively. It is also clear that there must be more user-friendly ways to
conduct a subject search. An interface that automatically searches the appropriate UDC code connected to the given index term would be of great help to the occasional user.

REFERENCES


2. ONNISTUUKO TIEDONHAKU - TIE TOKANNOJEN LAATU PUN TARISSA. KAHDEK SAN SUOMALAI SEN TIE TOKAN NAN ARVIOINTI JA KEY TETYT KRIT ERIT. (in Finnish)(Any luck with the search - the evaluation of eight Finnish databases), Tietopalveluseura, Projektiraportti, Espoo, 1991: p. 73.


