John Catlow
Mirosław Górny
Rafał Lewandowski
Institute of Linguistics
Information Systems Group
Adam Mickiewicz University, Poznań, Poland

Keyword binding as a method of reducing the length of indexes in library catalogues (based on the experience of Digital Library of Wielkopolska)

Abstract

This paper presents a relatively simple and cheap method for shortening the subject indexes in library catalogs. The method involves taking a set of several dozen general concepts, characterized by a low semantic awareness barrier. Built around these words are subindexes made up of the words which appear in descriptions containing a particular general concept. The effectiveness of the method was studied by analyzing the content of fragments of subject indexes of the NUKAT central catalog of Polish libraries, the University Library in Poznań and the Library of Congress. Compared with the subject headings language method, this method reduces the length of an index by an average of two-thirds, and makes it significantly easier for readers to navigate the vocabulary used by the cataloger. This method has been developed for the needs of Digital Library of Wielkopolska, and will probably be used in all regional digital libraries in Poland.

Introduction

When searching databases containing bibliographic descriptions, a user is sometimes compelled to browse indexes of the words used in the descriptions. This is usually due to the need to determine what range of vocabulary was used by the catalogers.

A reader very often does not know what words have been approved as appropriate to the catalog. For example, a reader searching for the term *rakiety przeciwlotnicze* ("anti-aircraft rockets") will not find it in the KABA dictionary (the dictionary of subject headings used by Polish academic libraries), even though the

term is commonly used both in the specialist literature and in journalism (it returns more than 4000 Google hits). However searching for the same phrase in the subject heading field using a search engine which performs parallel searching of the catalogs of Polish academic libraries (the Distributed Catalog of Polish Libraries or KaRo, http://karo.umk.pl/Karo/) shows that the term has never been used in the catalogs (a negative response was returned from 33 Polish university, technical university and specialist humanities libraries, as well as from the NUKAT central catalog of Polish libraries).

Similar results were obtained from an experiment with the term *armaty przeciwlotnicze* ("anti-aircraft cannon"), which was not found in the catalogs of 42 libraries, even though it returns 3680 hits on Google.

This does not mean that the catalogs do not contain any terms relating to the subjects in which the reader is interested. There exists the term *technika rakietowa* ("rocket technology") under which one can find items relating to anti-aircraft rockets. There is also the term *działa przeciwlotnicze* ("anti-aircraft guns") which is equivalent to the term *armaty przeciwlotnicze*. However, in order to find this out, it was necessary to browse the dictionary.

If readers fail to find the terms they seek immediately, they will try to think of other suitable terms, or else will be forced to browse the whole of the subject index of the catalog of the library in question.

This task is not especially onerous as long as the indexes remain relatively small. In most cases, however, the size of these lists means that browsing them requires a significant amount of time. This is particularly so in the case of collections covering many fields, described using vaster and more varied vocabulary than in the case of monothematic collections.

For example, the list of keywords used in the Digital Library of Wielkopolska[1] contains approximately 20,000 objects. This was created as a result of the cataloging of around 50,000 items. Browsing such a long list would be a burdensome task.

The word lists used in large library catalogs, and particularly in central catalogs, are obviously many times longer.

Admittedly the reader usually does not have to browse the complete list, only a selected fragment of it. Nonetheless it takes a significant time to browse a list of even

[1] *Wielkopolska Biblioteka Cyfrowa* (Digital Library of Wielkopolska), http://www.wbc.poznan.pl/dlibra

a few hundred entries. The only solution is to strive to make the index fragments which a user has to browse as short as possible. This is the principle behind the method used at the Digital Library of Wielkopolska.

1. The semantic awareness barrier

The keyword binding method is based on a certain way of organizing the indexes which are presented to a user of the system. Normally the user gets an index containing all of the keywords which are used in the system. (Naturally the words used in each field of the bibliographic description may form separate indexes. However we are interested here only in the index of keywords.) Now we present the user with not one index, but a certain number of shorter indexes. The global index has been divided into shorter fragments. In this way browsing becomes faster.

But according to what rule is the index to be divided? The cheapest and fastest method appears to be to generate, using a computer, lists of keywords which have some common feature. In this case the feature is a specific term which appears in the keyword field of bibliographic descriptions along with other words of the subindex. Namely, a given subindex is formed by all keywords selected from those keyword fields of a bibliographic description in which a given specific term **always** appears.

It is clearly important, however, that these specific terms be selected appropriately. A key feature of the chosen terms should be that they present a "low semantic awareness barrier".

This barrier is one of the key problems in the information searching process. It is due to differences between the semantic awareness of the user of information and that of the cataloger, i.e. the creator of the metadata used by the information system. In other words, the user of the system often understands certain concepts differently than the cataloger does.

Search failures are largely a result of this difference. There can therefore be no doubt that information search methods ought to take account of this problem and attempt to minimize the difference. This does not always happen, because sometimes the creators of search methods fail to appreciate the importance of this problem or remain unaware of it. It is also the case that there is not always a satisfactory way of solving the problem.

We can, of course, imagine an information search system created by persons whose semantic awareness is identical, or at least similar, to that of users. This

situation might be treated as one of the ways of breaking the semantic barrier. If, for example, the user of an information system and the cataloger have had similar education, dealt with similar problems and had similar experience in the field to which the information search system relates, it can be expected that they will "understand each other" relatively well. However, achieving such "semantic agreement" is exceptionally difficult. It is doubtless possible only in respect of small groups of users and catalogers. Moreover, the "semantic cohesion" of the two groups would deteriorate with time, under the influence of various factors.

Unfortunately, in the majority of cases – particularly with systems covering multiple fields – the level of the semantic barrier is extremely high. This is particularly the case when users have highly differentiated sets of knowledge and experience. This is the case, for example, with Internet-based systems, of which the Digital Library of Wielkopolska is an example.

Terms for which the level of the semantic awareness barrier is low include **sufficiently general terms** and **sufficiently narrow terms**.

The use of such terms as subindex headings means that a user who selects a given subindex based on a corresponding heading is highly likely to get to the index fragment most appropriate to his or her query.

General concepts are characterized by a large volume of meaning. Therefore, in spite of evident differences in the way they are understood by individual users, at a certain level of generality it is normally possible to determine a meaning which is common to all users. Then all individual ways of understanding some term (within a certain cultural milieu, obviously) fall within the meaning of the term at a high level of generality.

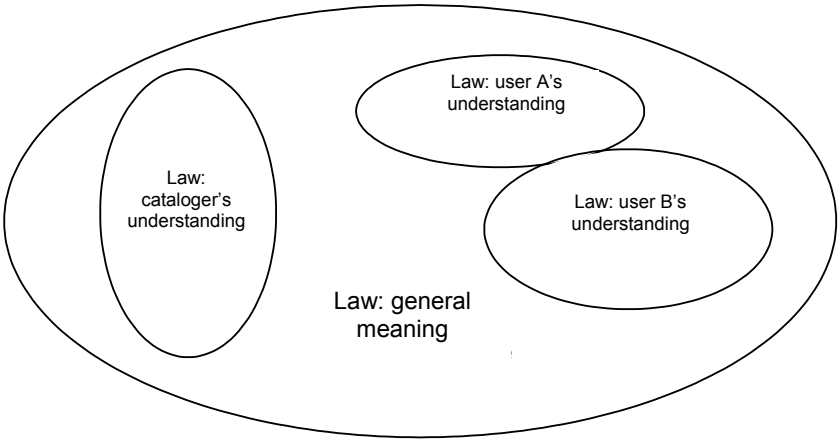

Diagram illustrating the dependences between the semantic awareness of those involved in the information process in relation to terms with a high level of generality (characterized by large semantic volume).

The second group of terms for which the level of the semantic awareness barrier is low consists of **terms of very narrow meaning**, which for this reason are universally understood in an unambiguous way (again we are considering a cultural milieu which is uniform in many respects). Terms of narrow meaning include, for example, some proper names, specialist terms, names of rare fauna and flora, etc. Naturally when using terms from these categories we create an information barrier to users whose knowledge is more limited.

2. Comparison of the keyword binding method with the subject heading language method.

a) The subject heading language method

Use of the subject heading language method is, in a certain sense, a similar solution to the keyword binding method. However this similarity is limited. In an expression of the subject-heading language the main heading stands first, with defining words ("subdivisions") associated with it. An index is a list of word associations.

The advantage of this solution is high accuracy of response. The defining words limit the set of publications returned to those which may interest the reader. If the reader succeeds in finding a suitable association of words, he is highly likely to accept that he has found a publication on exclusively the subject which interests him. Sometimes, however, he suspects that not all publications of significance for him are described with the words used in the construction of the query. Some of these words he will no doubt find in the bibliographical descriptions of the items returned. However he would like to be sure that he has found everything, and will no doubt attempt to browse other parts of the index.

If subject heading language is used in the cataloging of a relatively large and thematically differentiated collection of publications, the result will be a relatively large subject index.

This is a result of the many possible ways in which a main heading can be associated with subdivisions. If the number of these subdivisions is relatively large, it is fairly obvious that we may obtain a very long list of headings in the index. For example, if we use three types of defining words, the numbers of words of the respective types being o_1, o_2 and o_3, then the length of the list obtained will be equal to the product $o_1 \times o_2 \times o_3$.

For example, if we have 10 geographic subdivisions, 10 chronological subdivisions and 5 form subdivisions (representing the bibliographic, literary or artistic form in which the material is organized or presented), then the number of possible combinations is 500.

Naturally it is unlikely that the cataloger will have used all of these possibilities. Firstly, the collection may not contain publications corresponding to all combinations. Secondly, certain combinations may for various reasons be considered

impossible. Thirdly, a cataloger does not always make use of all permissible subdivisions.

In spite of this, the number of combinations is usually still very large. It is this which is the main defect of this solution.

Let us consider an example from the NUKAT catalog in 2007/2008[2]. A reader wishes to check which words have been used in the cataloging of works in the field of chemistry. She submits a query containing the word *chemia* ("chemistry"). This returns an alphabetical list of 3018 multiple-word phrases in which the word *chemia* appears not only as a main heading, but also as a topical subdivision. This makes it necessary to browse the content of approximately 300 pages (the system displays ten items per page).

The same list displayed as a list of single headings contains 1156 items[3]. Therefore this will be the number of words in the subindex created by the method of binding keywords with the heading *chemia*.

If we submit the query *prawo* ("law") to the catalog of the library of Adam Mickiewicz University, Poznań (which also uses a system of subject headings), we are returned a list of 2408 multiple-word phrases. When broken down into single words, this list is shortened to 881 words[4].

Similarly, the word *historia* ("history") in this catalog returns a list of phrases consisting of 733 word combinations. However the number of single words is only 342.

The word *chemia* in the catalog of the library of Adam Mickiewicz University returns 296 phrases. The list of single words consists of 193 items.

In the Library of Congress catalog the word "chemistry" returns 1564 phrases. On decomposition this becomes a set of 561 single words[5].

---

[2] From the authors' own research using the NUKAT catalog. NUKAT is a union catalog for Polish academic and research libraries. It is built by means of shared cataloging, which means that each description of a document is built only once, stored in the NUKAT database, and downloaded as needed to the local catalogs. Over 1100 librarians from 81 libraries contribute to NUKAT. The NUKAT database contains bibliographic records and authority records. All records are currently entered online. Moreover, NUKAT contains all bibliographic records from the former Union Catalog of Serials and all authority records from the former Union Authority File. (http://www.nukat.edu.pl/)

[3] Single-element phrases made up 7.59% of the total, two-element phrases 37.57%, three-element phrases 35.82%, four-element phrases 15.18%, five-element phrases 3.51%, and six-element phrases 0.33%.

[4] Single-element phrases made up 17.36% of the total, two-element phrases 29.82%, three-element phrases 31.60%, four-element phrases 15.78%, five-element phrases 5.23%, and six-element phrases 0.21%.

[5] Single-element phrases made up 4.6% of the whole, two-element phrases 34.72%, three-element phrases 42.03%, four-element phrases 13.30%, five-element phrases 3.39%, six-element phrases

b) The keyword binding method in practice

It is necessary to determine a list of keywords with a high level of generality in relation to the subject matter covered by the database. It is not easy to lay down criteria for the choice of such words. They depend on the number and size of the independent thematic domains represented in the database. However it appears that account ought to be taken of the semantic awareness of the **average** user. This implies a set of words which are in common use, such as "chemistry", "physics", "mathematics", "law", "economics", "history", etc. The semantic volume of such keywords is very large. Even if the cataloger and the user understand the content of a given general keyword differently, their understandings are nonetheless contained within the area of meaning of that word.

The number of these keywords is largely dependent on a desire to obtain a reasonable index length which is acceptable to users. Another factor of some significance is the influence of the number of these keywords on the efficiency of the work of the cataloger. It is certainly easier to operate with a collection consisting of only a few dozen keywords, if only because the cataloger will succeed in learning them more quickly. Extension of the list of such keywords to a few hundred items, for example, clearly makes both searching of the index and the cataloging process itself more difficult.

We adopt the assumption that all bibliographic records must be denoted with at least one general keyword, although the possible number of general keywords is unlimited.

A reader wishing to browse an index selects one main heading (e.g. from a relevant list). She receives a list of the headings which are used in descriptions in combination with the selected heading (i.e. are bound to the main heading). These form thematic collections (lists) which are short enough for the user to be able to browse them in a reasonable time. From the list thus obtained, the user can select appropriate terms for use in building a query.

We estimate that for 20,000 entries in a dictionary, at least 200 general headings are needed. This means that under each general heading there will be approximately 100 words, equivalent to, for example, 10 pages to browse. In reality, however, it is highly unlikely that such an even distribution would be obtained.

0.90%, and seven-element phrases 0.06%.

This distribution will be shaped by policy followed as regards collection (certain publications may significantly dominate, because of requirements) and annotation (certain publications are annotated in more detail than others). In this way it is possible, acting consciously, to obtain any possible form of distribution. However the most probable will be one given by general laws of distribution. This means that approximately 20% of general headings (about 40 headings in our example) will associate with approximately 80% of words (about 16,000 in our example.). This means that certain general headings may be bound with about 400 words each.

Thus in effect these 40 headings (which may well also be among the most commonly used) ought to be enhanced by the addition of narrower headings. For example, “Law” as a main heading may be divided into “Administrative law”, “Criminal law”, etc.

Comparison with the catalogs of large Polish university libraries (with more than one million volumes) indicates that the subindexes created as a result of keyword binding are on average three times shorter than the lists obtained by searching for a general heading in catalogs using the subject heading language method.

Naturally the inclusion of general terms in bibliographic descriptions requires conscious action on the part of the cataloger. Should the subindexes begin to extend beyond a sensible length, it will be necessary to undertake the time-consuming task of introducing narrower general terms into the descriptions (recataloging). If the excessive growth of the indexes is anticipated, then naturally these terms can be introduced earlier. A type of systematic catalog will then be produced. It should be remembered, however, that action of this sort is costly and labor-intensive. The introduction of ever narrower terms requires more cataloging time, in view of the need to spend more time analyzing the content of the publications being cataloged. Moreover there must be made available to the reader a transparent schema presenting the hierarchy of the terms used. As a result we begin to build the framework of a systematic catalog.

However, at the same time, it is possible to create a similar system by generating the subindexes automatically. A computer program can select the terms most commonly appearing in the catalog and set those as subindex headings. The result will be lists of words which are used in the bibliographic descriptions in combination with a commonly appearing word. The words most commonly used in catalogs are generally words with a low semantic awareness barrier.

Clearly it is possible to build a subindex in which the header (binding element) can be any desired word from the catalog. However, there are technical limitations. Such an index would have to be generated in real time, whenever requested by a user, which would place a significant load on the server. It would obviously be quite impossible to generate indexes for all words every day, as the servers would not be able to cope with such a task.

In other cases – those described above – the index is generated (updated) once every 24 hours.

Summary

The new method makes it easier for readers to find their way around the content of subject indexes. Firstly, a user can browse word lists which are several times shorter, and secondly, the lists contain single words rather than the word combinations offered by the subject heading language method. This significantly shortens the browsing time, since not only does the list contain fewer items, but it also provides much easier perception. It is also of significance that the system does not require the cataloger to build multiple-element phrases: in the bibliographical description it is enough to use single words[6]. This reduces the time needed for cataloging.

No investigation has yet been carried out into the degree to which this method makes the use of a controlled dictionary unnecessary. However, we suspect that shortening of the index enables the reader to browse all words used in a given thematic domain (words concentrated around a single thematic concept). This may mean that a controlled dictionary will not be needed, particularly because synonyms are widely used in the system.

This method might be improved through the use of narrower general terms or the creation of subindexes for any desired words. Unfortunately in both cases the costs

[6] Admittedly there may then arise false associations which do not occur with the subject headings method. For example the use in the same description of the words London, architecture, 20th century, Great Britain and 19th century may falsely suggest that the work concerns 20th-century London architecture when in fact it happens to refer only to London architecture of the 19th century. However associations of this type occur very rarely and do not present a major problem to readers of digital libraries, because a reader can, after finding an interesting description, immediately verify the content of the work by opening it. Such errors were significant in times when books needed to be borrowed. Then the reader and library wasted time unnecessarily as the inappropriate item was retrieved from the store.

greatly exceed the benefits to the user. It can therefore be expected that in practice it will not be profitable to make further improvements to this method.

Searching modern information systems cannot be done based on one single method. Depending on needs, it seems appropriate to apply combinations of methods. An important issue is the ability to make a comprehensive assessment of the effectiveness of searching. An assessment method is sought which would make it possible to select an optimum set of methods for organizing a collection, taking account of both the quality of search results and the costs incurred by the institution maintaining the information system, as well as the time taken by users to search for information.