Enriched thesauri
and their uses in information retrieval and storage

Discussion paper

Michiel Hazewinkel
CWI, Amsterdam
POBox 94079, 1090GB Amsterdam
mich@cwi.nl

Introduction.

The subject of this discussion paper is information storage and, especially, information retrieval from large and very large collections of objects. The focus is on scientific objects such as papers, tables, programs, handbooks, manuals, ... All of these will be referred to here and below as documents. They can be quite heterogeneous in form and content and the storage medium can be varied, though it is assumed that at least adequate metadata (see below for this concept) are, or will be, attached in machine readable form.

Quite apart from the sheer size of the problem there are a variety of reasons (for instance linguistic ones having to do with morphological variations, synonyms, and homonyms) that indicate—personally I would put it much stronger—that full text search is not a real alternative. This has always been the major reason to work with a "controlled vocabulary", here interpreted as a standard list of key phrases for a given section of science and/or technology, [2].

Another main reason is multilinguality. It is feasible to have essentially the same thesaurus (of key PHRASES (possibly with additional identifiers to deal with ambiguities)) in several languages with good explicit correspondences. It is in any case far simpler to realize such a thing than to do anything like automatic translation.

2. Metadata. The key to dealing with (very) large collections of scientific documents is metadata. This concept includes such things as bibliographic data such as CIP data and Library of Congress cataloguing data; it also includes classifications in terms of one of more (standard) classification schemes and attaching a number of key phrases to a document. The importance of metadata is illustrated, e.g., by the very substantial effort that Elsevier puts continuously into maintaining the thesaurus behind EMBASE, the database of Excerpta Medica, [4], and the even more considerable effort involved in adding adequate metadata in the form of key phrases and words from that thesaurus to each document of EMBASE/Excerpta Medica.

The question of metadata rather naturally divides into two interrelated parts:

- Questions of design and implementation of a framework, or container, for metadata, including how to translate one metadata scheme to another.
- Questions concerning what should be put into the various slots of the container, including how these data should be used for information retrieval purposes and how to generate metadata for a given document (semi-)automatically.

For instance the Dublin Core famework (named for Dublin, Ohio, rather than Dublin, Ireland) consists of the elements: Subject, Title, Author, Publisher, Other agent, Date, Object type, Form, Identifier, Relation, Source, Language, Coverage. Most of these elements can be replicated as often as desired or necessary, [10]. Further work, extending the Dublin core, on developing containers for Metadata is in progress under the name Warwick framework , [5]. This is led by NCSRTL, has the participation of Library of Congress and the NSF/ARPA/NASA digital library initiative and involves the national libraries of the UK, Norway, Finland, The Netherlands, Australia, New Zealand, and OCLC, NCSA, etc.

For most of these elements what kind of information should be put into them is fairly clear. A clear exception is the element *subject*, which encompasses such things as classifications according to various classification schemes and key phrases describing the subject matter of the document. In addition to the standardly used bibliographic data such as title and author (or identifier) this is the main source of data for information and document retrieval at the metadata level.

Much of the what follows below is adressed specifically to the matter of searching in quite large and very large collections. Whether these collections are distributed over a network or are all in the same place is largely irrelevant for the issues addressed here. Experienced users of the various search engines on the Web (WWW) will know how difficult it is to keep down the number of hits (while keeping them relevant) without sophisticated search tools. The problem is widespread and generally recognized.

3. Enriched thesaurus. Definition.

An enriched thesaurus for a given domain in science and/or technology consists of several components.

(i) The main and central component: a list *L* of key phrases sufficiently rich to describe the domain in question adequately. For instance for mathematics I estimate that *L* should have about 120 000 entries. For computer science about 90 000 should be adequate.The Encyclopaedia of Mathematics, [6], has about 30 000 entries in its index volume (not counting inversions and linguistic morphological variations). The thesaurus behind EMBASE and Excerpta Medica works with about 25 000 basic entries.

The key phrases in *L* should be long enough to carry real information (single words in science only rarely do so). All entries in *L* are prepositional noun phrases (PNP's). A prepositional noun phrase consists of noun phrases (NP's) (strings of adjectives and nouns with or without particles, linked together, possibly iteratively, by prepositions (such as 'of', 'for', 'on'). For example: "differential invariants of four dimensional manifolds".

(ii) A distance function on *L* which reflects how related items in *L* are and how likely they are to occur together in the same documents.

(Technical note. Storing a distance function on a set of this size explicitely, takes a horrible amount of storage space; fortunately there are far more efficient ways of doing this, [7].)

(iii) A partial order on the set *L*, indicating which phrases (concepts) are more general (broader terms) and which are less general (narrower terms).

(iv) A labeling of each item of *L* with a node (most often a leaf) of one or more classification schemes for the domain of inquiry in question, such as the PAC classification scheme for Physics and Astronomy, the INSPEC classification scheme for electrical engineering, computers and control, the MSCS (Mathematics Subject Classification Scheme) developed by the American Mathematical Society and FIZ/STN (Zentralblatt für Mathematik und Grenzgebiete).

Components (ii) and (iii), respectively, refine and replace the classical thesaurus notions of 'related terms' and 'more general term (broader term)', 'less general term (narrower term)'.

Note that the classification schemes used in component four, which are also part of the enriched thesaurus, also get enriched themselves in that each node acquires a list of key phrases carrying that node as classfication thus giving that node content and meaning far beyond the short description given by its name. Such descriptions are a valuable tool for researchers and others when assigning classifications to a document.

4.. Some background: controlled lists and error correcting codes.

The idea of using only part of all objects of a given kind, usually a quite small part (relative to the total) but a quite well distributed one, for purposes of description, understanding, manipulation, ... is one that occurs with frequency in various parts of science and technology. Thus one has for example (truncated) singular value decomposition in mathematics and signal processing, principal component analysis in econometrics (economics), factor analysis in psychology. These are mathematically basically the same and can also be used for intelligent information retrieval, [3]. Another instance is lexical stemming in linguistics and still another is coding theory in communication and information theory. In all cases the idea is to systematically get rid of such things as noise or contamination, unwanted variations, less important or unimportant factors.

The idea of a standard (but dynamically evolving) thesaurus is perhaps best thought of as a linguistic variant (and information retrieval variant) of an error correcting code.

5. Semi-automatic generation of an enriched thesaurus.

Take a large enough corpus and divide it into suitable chunks called documents. For instance take the 700 000 abstracts of articles in the STN/FIZ database Math (ZMG data), or take as documents the sections or pages of a large handbook or encyclopaedia such as the Handbook of Theoretical Computer Science, [9], or the Encyclopaedia of Mathematics, [6]. Now use a parser for prepositional noun phrases (PNP's) (or an automaton recognizing PNP's) or a software indexing program, to generate from these documents a list of key phrases, keeping track of what phrases come from what document. This gives component (i) of an enriched thesaurus and, using e.g. Hamming distance, the incidence matrix between documents and key phrases thus created also provides component (ii).

(Technical notes. For the results of a first exercise in this direction see [1, 8]. Such automata and/or indexing software packages exist in several varieties. The Hamming distance between two terms is the number of documents which contain the one term and not the other; more sophisticated variants, embodying the idea of weights (reflecting importance, length, ...) are obvious.)

To generate component (iii) (less essential in the present context and in view of the uses to be made of the enriched thesaurus; see below) one can generate the network of balls defined by the the discrete metric space of phrases defined by components (i) and (ii), [7].

(Technical note: obviously components (iii) and (iv) of an enriched thesaurus should be compatible in a suitable way; this can be done by applying a suitable clustering method to the discrete metric space of key phrases (thus yielding a bottomup classfication scheme). The general problem of finding a best partial order compatible with one or more classification schemes is largely open.)

Generating component (iv) of an enriched thesaurus is relatively staightforward if one works with a corpus like the ZMG data or a similar database maintained by STN/FIZ for theoretical computer science (to which adequate access has been arranged). The documents in these corpora are labelled with classification codes from the corresponding classification scheme. Similär access to the ACM and INSPEC (IEEE) databases will be rather necessary to do a good job for all of computer science.

5. Thesauri rather than one thesaurus. The title of this essay speaks of thesauri rather than one thesaurus. In my opinion, it is flatly impossible to generate all at once an adequate thesaurus for all of science or even of all of a largish field like physics. The field of mathematics is about as large as can be handled at once I think. The solution is to have several thesauri which may overlap and be of different levels of detail. Much like in a geographical atlas one has maps and charts which overlap and can have different scales (and be of different kinds).

(Technical note. There are intersting and open mathematical and algorithmic problems in obtaining maximally good overlap correspondences.)

6. Searching. Once one has an enriched thesaurus, in the sense defined above, in place there become available advanced intelligent search possibilities based on the contents of the subject fields, classification fields, title field, key word fields, and abstract field in the metadata container attached to each document in the collection at hand. Note that in the Dublin core design for metadata all these fields, except the title field, are replicas of "subject". These search possibilties can be seen as being of two kinds:

Direct thesaurus mediated search

Dialog mediated search.

Both come with a 'local search' refinement.

For the first kind the searcher types in a Boolean combination of items in the thesaurus list, the author list, the classification scheme lists, and obtains as response the number of hits. The answers thus obtained can be Booleanly combined with the results of other queries and the final result can be displayed as a list of documents.

For the second type of search the searcher types an arbitrary potential key phrase (or several). A *PNP comparer* answers with a list of the *n* closest phrases which actually occur in the thesaurus and asks the searcher which ones he would like to use in some Boolean combination. Here *n* is a number that can be specified by the searcher with a default value of for example 5.

(Technical note. A PNP comparer in the present context is very much like a decoder in coding theory. The general problem of defining an adequate distance on sentences for a natural language grammar which allows for the automatic correction of small grammatical mistakes is a long standing and much neglected open problem; the same problem is at hand here but only for the very simply structured class of PNP's and several admittedly ad hoc solutions suggest themselves.)

For the local search refinement remember that the space of key phrases comes with a metric (and so do the spaces of classification scheme items). Thus the searcher can specify a centre, which is a set of such items, and request a seach for specified other items which are within a searcher determined distance of the given centre.

For a final search possibility note that the documents in the total distributed repository have key phrases and classifications attached to them. The resulting incidence matrix also defines a metric on the space of documents in the total repository. Thus it is possible to specify a document centred search which can be of one of the above types or can even be full text if so desired.

7. Updating issues.

Science fields evolve. To remain useful an enriched thesaurus must be regularly updated. Also, certainly, besides using key phrases from the existing thesaurus, authors must be free to invent ones of their own and put them into the subject fields of the metadata container.

(Technical note. Should the key phrases in the subject fields in the Dublin core be marked depending on whether they come from a specified thesaurus or were freely invented?)

Dynamic updating of the classification schemes used ususally, of course, will be outside the scope of a group engaged in constructing a thesaurus for a given field. With this restriction, updating an enriched thesaurus as defined above is a relatively straightforward matter. It does require, however, a small *new key phrase detector*, which lists the new key phrases invented by the authors which contribute documents to the total repository. These become candidate new key phrases which can be incorporated, if found suitable, at some future time.

8. Multilinguality.

As suggested above automatic translation of a query in one language to one in another one is difficult to realize directly. However, it is possible, given adequate corpora, to do the 'thesaurus' generating job described above for several different languages. The resulting metric spaces should be more or less similar and that can be used to establish reliable correspondences between them. More precisely, using existing dictionaries, the metric space structures can be used to provide suggestions for translations and to provide checks on the adequacy of the translations. Again it needs to be stressed that it is translations of KEY PHRASES rather than individual words which are needed

9. Atlas of science and technology.

As said above, I do not think that a job such as was described above can be done all at once for all of science, not even for all of a largish established and well defined discipline like physics. The aim is to generate thesauri (in a similar way) for many partially overlapping parts of science, and, where possible, to do it at varying levels of detail. Thus there will be many thesauri in varying levels of detail for varying parts of science and addressing different kinds of information. The whole can be likened to a geographical atlas with its many different charts at different scales and of different kinds (geologica, demographical, linguistic, mineralogical, political, ...). Just as in the case of such an atlas the correspondences (which in the case of a geographical atlas are rather obvious) need to be established. Here again the metric structures of the specific thesauri which make up the atlas, can play a guiding role.

10. Concluding remark.

The above is not the description of an existing thing. It is instead an outline of a possible project, that, in my opinion, can be carried out with basically existing tools, and that should be carried out.

References.

1. *Subject index volumes 101 - 150*, Theoretical Computer Science **150**:2 (1995), 197-313.

2. Jean Aitchison, Alan Gilchrist, *Thesaurus construction*, Aslib, 2-nd Edition, 1990.

3. Michael W Berry, Susan T Dumais, Gavin W O'Brien, *Using linear algebra for intelligent information retrieval*, SIAM Review **37**:4 (1995), 573-595.

4. Ian Crowlesmith, *Creating a treasure trove of words*, Elsevier Science World. 14-15, 1993.

5. Juha Hakala, Ole Husby, Traugott Koch, *Warwick framework and Dublin core set provide a comprehensive infrastructure for network resource description*, html document, 1996. URL: http://www.bibsys.no/warwick.html

6. M Hazewinkel (ed.), *Encyclopaedia of mathematics; 10 volumes*, KAP, 1988-1994.

7. Michiel Hazewinkel, *Classification in mathematics, discrete metric spaces, and approximation by trees*, Nieuw Archief voor Wiskunde **13** (1995), 325-361.

8. M Hazewinkel, *Preface to the index for volumes 101-150 of TCS*, Theoretical Computer Science **150**:2 (1995), 195-197.

9. Jan van Leeuwen (ed.), *Handbook of theoretical computer science*, Elsevier, 1990.

10. Stuart Weibel, Jean Godby, Eric Miller, *OCLC/NCSA metadata workshop report*, OCLC, 1995. URL: http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.

Saturday, 4 May 1966, Revised: Wednesday, 05 February 1997