Automatic keywording of High Energy Physics

David Dallman, Jean-Yves Le Meur

CERN, Geneva, Switzerland
5 October 1999

Bibliographic databases were developed from the traditional library card catalogue in order to enable users to access library documents via various types of bibliographic information, such as title, author, series or conference date. In addition these catalogues sometimes contained some form of indexation by subject, such as the Universal (or Dewey) Decimal Classification used for books. With the introduction of the eprint archives, set up by the High Energy Physics (HEP) Community in the early 90s, huge collections of documents in several fields have been made available on the World Wide Web. These developments however have not yet been followed up from a keywording point of view.

We will see in this paper how important it is to attribute keywords to all documents in the area of HEP Grey Literature. As libraries are facing a future with less and less manpower available and more and more documents, we will explore the possibility of being helped by automatic classification software. We will specifically mention a project being carried out at CERN (European Laboratory for Particle Physics) for testing this automatic keywording.

SEARCHING DOCUMENTS BY SUBJECT

There are two main uses of a bibliographic database. The first one is to search for a specific item which one already knows about, and wants to find out if the library has it, and if so, to get access to the document. This is the so-called referral approach, a bit like looking up a piece of information in an encyclopaedia. You know it is there, you just need the answer.

The other main use is when one has a specific problem in mind, and wants to find documents which address that problem. It is only with this second type of use that we are concerned here. Basically, this means a subject-based approach to the library collection.

1) Subject connections via references

There is already a system of searching academic literature in a thematic way without any kind of intermediate database. This is via the references to other work which have been an accepted and important part of scholarly publication since the very beginning.
Starting from a core document, one can gradually widen the scope using the references and hopefully arrive at some fairly complete set of relevant documents. The electronic age has again enhanced such an approach without however changing it in principle. References in an electronic document can be links to the electronic versions of the documents referred to.

The main obvious drawback in this approach is that authors may not have referred to all the relevant material, either due to deliberate omission or just because they do not know about it.

Another possible disadvantage is that by definition one can only refer to what already exists at the time of writing a document ! From the core document, past documents will be reached but all new documents will be missed. However, this could theoretically be solved through a database of such references, by forward searching from a document to retrieve all other documents which have referred to it later.

In practice, the connections via references is not an adequate approach for users needing for an exhaustive list of available documents related to a given topic. It takes too long and the full coverage is not guaranteed. The other solution - querying directly a bibliographic database - is much faster but it may still result in an incomplete result.

2) Recall and precision in searching

In evaluating the value of search results, two different concepts are of great importance: the first is the recall factor and the second is the precision.

The recall factor measures how many of the total number of documents which should be retrieved by the search, are in fact retrieved. Some relevant documents may be missed by a given search strategy. Taken on its own, this factor should obviously be as close to 100% as possible.

*recall = 100 * number of documents retrieved / total number of relevant documents*

The precision measures what part of the documents retrieved actually belong to the desired sample, the rest being undesired documents which have somehow also managed to satisfy the search criteria. Again, when considered alone, this should also be as close to 100% as possible.

*precision = 100 * number of relevant documents retrieved / number of documents retrieved*

The problem is that these two measures of search efficiency are not independent, in fact there is an anti-correlation between them. If you try to get the recall factor as high as possible by using a more complex search strategy, you will also tend to pick up more "background" documents which you do not want. Conversely, if you want all your retrieved documents to be relevant, you will have to pay the price of missing quite a lot of relevant documents too.

Database search engines may offer some features which are designed to improve the precision of the search. Words can be strung together as a phrase, and this phrase searched for. However searching for a phrase of more than three words is likely to result in a low recall factor, because of the flexibility of natural language (particularly English) in representing nuances of meaning by variations in word order.
In another approach, limits can be placed on the maximum number of intervening words which are allowed to occur between a pair of chosen words (proximity searches). The CERN Library database has such a functionality but it is rather cumbersome to use at present.

Of course, it is not just the search strategy which counts, but the result of the search strategy when applied to the data. Therefore, exactly which data for a given document are available for searching influences the search result.

3) Data from the document itself which are available for searching

Here we are only really concerned with data concerning the subject matter of the document, so things like author names do not play a role. Of course, searching for an author can result in retrieving a certain subject, but it is almost never the case that this author is involved with all the documents in that subject area.

Traditionally, the title is the item of bibliographic information which expresses the subject content. However, a title is usually far too short to contain a complete description of the subject area in a way which can be used efficiently by a search engine. A specialist reading the title may understand what the document is about, but he is using all sorts of prior knowledge into which context he plugs the new title. Therefore the recall factor of a title-based search is likely to be low. Furthermore, as the number of documents in the database steadily increases with time, the precision of title-based searches is likely to decrease as well.

An extension of this, which has become much easier to realise for electronic documents, is that more of the text than just the title can be used for searching. In particular, extending the search to the abstract is a very useful step. This has been done with the CERN HEP database since we started handling electronic preprints in 1994. However it has to be remembered that all words in the text are treated equally, so a mention in an abstract of a term by way of contrast and not because it is dealt with in the document, will still cause it to be indexed.

In principle, the full text of the document could be used for searching, but in practice this has not been done for documents in the HEP field. Considering the huge number of documents produced in this field, searching on full text would probably give a good recall factor but the precision would be far too low to be really useful. This is surely why we do not know of any project for indexing the whole text of HEP literature.

An alternative approach is to supply additional data concerning the subject material, and to use this for searching.

WHY DO AUTOMATIC KEYWORDING ?

Adding of subject material is called subject indexation or keyword enhancement. When we say "keyword" it could of course be a phrase of two or more words. There are two very different ways of doing this : to choose terms from a fixed thesaurus or to use free keywords which can be chosen by the indexer at will. The strategy of assigning keywords will obviously depend on which parts of the document itself (title, abstract, full document) are also available for searching.

1) Adding data to the documents

a/free keywords

Allocating keywords on a free basis could also use terms which are not present in the document, but in practice this technique is mainly used for adding useful words or phrases taken from the text, such as section headings and other specific words which could help in improving the recall factor of the search. Free keywords can also be useful for indexing terms containing special characters which would not be completely recognised if they appeared in the title or abstract. For example, the CERN Library database normally breaks off indexing a word when it meets a non-alphanumeric character in a title or abstract, but it can be directed not to do this for a keyword field. Thus particles called W+ and W- would both be indexed as W in a title, but the full forms can be used as keywords and retrieved.

Free keywords can also be a useful way of adding synonyms of terms that appear in the text. But it would be better in general to handle synonyms at the search input end rather than adding them to each record when they occur.

b/ fixed thesaurus terms

The efficient allocation of keywords from a fixed thesaurus makes the most demands on the indexer, as the documents have to be well understood. The indexed terms may not appear in the same way in the text at all, which can give this method a big advantage over any strategy which just uses the text of the document.

Of course, such a method requires the existence of a complete, precise and up-to-date thesaurus, which is quite difficult to achieve in a rapidly-changing specialized research area like High Energy Physics.

2) Comparison between free keywords and fixed terms

In practice, these two forms of indexing are extreme cases. Real approaches have aspects of both, even though they may be closer to one than the other.

Thus, the drawback of having a fixed thesaurus is that the thesaurus itself has to be modified to keep up with developments in the field. This usually means that a new form of the thesaurus is issued at regular intervals, for example the DESY (Hamburg, Germany) HEP thesaurus has been updated every one or two years. Thus for searching back over many years, each time period should in principle be combined with the relevant thesaurus terms for that period. In practice, this complex procedure is rarely undertaken by the searcher. It could be built in as a front end to the search, but this has not been done yet for any of the databases in our field which use fixed thesauri.

On the other hand, free keywording can be chosen to conform to a minimum set of rules, instead of being completely free and just taking the words as they appear. For example, it could be decided to choose singular forms instead of plurals. In fact, after a period of use, listing the terms which have been given as free keywords does give a sort of "thesaurus in practice", which can then be used to standardize the keywords which are subsequently assigned, in order to improve consistency.

3) Influence of the keywording on search quality

Including the abstract instead of just the title in searching (with no additional keywording) increases the recall factor but probably reduces the precision. Use of proximity searching could offset this loss in precision somewhat. It is straightforward to measure the change in precision (within a database which permits searching in title and abstract separately) but the absolute recall factor cannot be measured like this, as one has no way of knowing which relevant documents have not been retrieved at all !

The free keyword system is designed to be used in conjunction with the other data like title and abstract. If used with the title alone, it probably improves both recall and precision. But it does not give much improvement over using title plus abstract, as free keywords are most of the time words already present in either the title or the abstract.

The fixed thesaurus approach aims at describing each document by a series of thesaurus terms in such a way that both the precision and recall are 100%. This aim might not be achieved in practice if the expertise of the indexers leaves something to be desired. It is very important to realise that searches should only be made using the thesaurus terms assigned, all other text like title and abstract should be ignored. Some kind of measure of how much the thesaurus keywording improves search results can be obtained by searching for the particular thesaurus term in the title or in the title plus abstract.

The table below shows the numbers of documents found in some examples. For the chosen term in the DESY HEP Index (which covers published HEP literature), we look for the occurrence of this term as a

keyword, then we look for its occurrence within titles in the same database. To compare with the abstract search, we use the CERN HEP database [1], whose coverage is similar.

We also look for the same term within the global scope of all eprints, published or not published. This gives an idea of the area which is not covered by the HEP Index.

Database:	**DESY**	**DESY**	**CERN**	**DESY**	**CERN**
Search performed / **Terms**	**By Keyword in *published* HEP**	**By Title in *published* HEP**	**By Title or Abstract in *published* HEP**	**By Title in *published and non published* HEP**	**By Title or Abstract in *published and non published* HEP**
"Supergravity"	***5945***	***1791***	***1000***	***2714***	***2908***
"Duality"	6257	1073	779	1557	2591
"Interface"	409	95	272	422	1005
"Bifurcation"	57	41	80	80	287
"Dielectric"	293	111	108	240	450
"Graphics"	147	9	41	140	161
"Measure"	**364**	**273**	**582**	**607**	**2556**

Comparison of searches by thesaurus terms and searches by title/abstract in the DESY and CERN databases (28/09/1999).

The differences in the results are striking. The most reliable numbers in terms of precision and recall are the ones in the first column.
This means that when a user finds "Measure" in 582 titles or abstracts of published HEP literature, only 364 of them only are really relevant to this topic. On the other hand, while a user may find the word "supergravity" 2714 times in the title, or 2908 times in the abstract of grey or published HEP literature, the number of documents actually relevant to this subject is more than double this. The other examples show the same kind of mismatch.

Moreover, it appears that the quantity of HEP literature without a classification (because it is not published) is quite large.

4) Conclusion

The added value of keywording based on a thesaurus is obvious, even when many other bibliographic fields are searchable. There is a direct relationship between the added value of keywording and the number of searchable documents: the more documents you keep, the more you need keywording.
A simple subject allocation cannot be satisfactory in the long term. Subjects need to be refined till they actually reach the precision of a thesaurus. The permanent increase of papers available in HEP will lead to a chaotic situation for Information Retrieval if a complete effective classification is not undertaken. In the next section, we will see what has been done so far in the High Energy Physics area.

Indexing by subject specialists is by the far the most precise method, but it is costly in terms of time and it requires highly-qualified people to do it. The question arises as to whether one could achieve a useful result by some automatic procedure based on the text of the title, the abstract or the full document.

TOWARDS AUTOMATION IN HEP

Before considering the automation itself, we give an overview of existing classifications in HEP. We describe the HEP specificity regarding the development of a keyword assigning expert system. Finally, we explain the tests that are currently being carried out at CERN.

1) Existing Classifications in High Energy Physics

Manual keywording has been carried out at DESY for more than 30 years. It covers all published articles in the various areas of HEP. The DESY HEP Index publication was the main output of this activity from 1963 to 1997. This publication itself then stopped but keywords are still allocated and they are searchable on the Web interfaces of the DESY [2] and SLAC [3] (Stanford, California) library catalogues.

A manual keywording activity used to be done at CERN as well. It started in 1983 with free keywording and was based on the HEP Index thesaurus (from 1989 to 1992). After this it was stopped due to lack of manpower.
Examples of fixed "commercial" thesauri are those used by INIS [4] (International Nuclear Information System, Vienna) and INSPEC [5] (Physics, Computing and Electrical Engineering Abstracts, UK). They are built manually and access is not free of charge. They are not sufficiently specialized in the HEP area and so are not really adequate for dealing with HEP literature.

Today, some articles do contain subject information supplied directly by the authors (usually only when the journal makes it a condition of publication!). So some journals have keywords, and quite a few journals have adopted the Physics and Astronomy Classification Scheme (PACS) classification [6] supported by the American Physical Society. However, these approaches are far from being complete, so they are not useful for global searching. Also the PACS classification is still too broad for detailed searching in a narrow field such as particle physics.

On the contrary, in the case of books, where the Universal or Dewey Decimal Classifications are widely used, this approach can be very useful for retrieving all books dealing with a particular subject. A Web interface enabling searchers to browse a partial UDC index exists for CERN Library book catalogue [7], HEP preprints and published articles have no such world-wide recognised classification.

The CERN project, in collaboration with DESY and SLAC, is to use an expert system for automatically deriving keywords and then to map them onto the DESY HEP Index. In other fields such projects have already been carried out rather successfully. The Medical National Center (MNC) [8] and NASA [9] use machine-aided indexing for example, to speed up their classification.

2) Particularities of HEP literature

Natural language contains a huge vocabulary and the syntax of languages is very complex. In traditional literature, a text can be processed by considering words as individual items. A dictionary of single words can be used as the basis for creating a knowledge base. In scientific literature, we consider that the meaning is expressed mainly through multi-word terms ("noun phrases").
In HEP, documents contain many particle symbols or equations which may be among the most relevant noun-phrases in the document. Describing the syntax of the sentences present in HEP literature requires at least the definition of a new type of word: the particle symbol.

In addition, the knowledge base needs to be set up differently for experimental and theoretical documents. It is also planned to handle another specific dictionary for technology-related papers.

Another particularity is the size of its electronic grey literature. It amounts to more than 100 000 documents since 1994 and is growing at the rate of about 20 000 per year.

3) Sokrates Learning System

SOKRATES [10] stands for "Self-organizing Object-oriented Keyterm Recognition And Text-Editing System". It derives from natural language key terms and keywords. Each new piece of information treated by the system is used to update a knowledge base. A learning system like Sokrates can be compared to a compiler: the input is a text written with a known syntax. The output is a condensed executable, like the set of key terms.

In terms of the earlier discussion, the Sokrates approach belongs to the free keywording type, where the free keywords must appear in the text and cannot be invented (except perhaps for synonyms if one chooses to build them in). So, for example, there is no way that this algorithm can return the term "Kac-Moody algebra" when the abstract says "graded Lie algebra", even though these are two names for the same thing.
The test of the software is divided into two parts: the derivation of the best key terms and their mapping onto the thesaurus.

a/ The Term Derivation

To run the extraction of the Key terms, three basic components are defined:

- ✓ A complete dictionary which is created and continuously updated. In this dictionary, individual words (any possible character type) are kept with the following two main attributes:
  - a code or type of word: "General", "Left", "Right", "Stop-word", "Particle", etc.
  - its frequency: number of times the word has been encountered in all documents processed so far.

- ✓ A knowledge base which stores all the key terms (single or multiple words) which have been selected together with their frequency.

- ✓ The rules for describing key terms (the "Text Description Language") which are expressed using the type of words. An example of a rule is:
  L A P G R … (Left, Stop Word, Particle, General, Right…), where L (R) enables one to specify that a word would only be significant when it appears to the left (right) of another relevant word.

An inference engine is able to match any rule to any text using the above rules and the dictionary of single words.

When dealing with a new document, Sokrates extracts all individual words and distinguishes old and new words. It counts repetitions and updates the dictionary. For new words (if any), it can ask an operator to provide the word characteristics (code).
In the next step, the term selector selects candidate terms. It uses the dictionary and the inference engine to extract all possible noun phrases. After this pass, a set of "valid" and "garbage" terms are available. The selector compares these derived terms to the established knowledge base, and keeps the ones for which a frequency threshold has been reached. The threshold can be defined differently according to the number of words in the noun phrases. For example, we could require that single words must have a frequency of 10, two-words terms a frequency of 5, etc.
If the number of derived keywords is too small, a third pass of the selector can be undertaken, with a reduced threshold definition.

b/ The Thesaurus Term Mapping

Different situations may occur:
- the key term exists in the thesaurus: the mapping is straightforward.
- the key term is similar to a thesaurus term: the correspondence can easily be established.
- the key term does not exist in the thesaurus: a subject indexer could associate one or more terms from the thesaurus to each key term found. This would only need to be done once for the whole dictionary; it would remain valid for all incoming documents. Only new key terms would need to be associated at later stage. Thus the savings over indexing on a document-by-document basis could be considerable.

4) CERN test: the status

The data for the test consists of about 1400 abstracts of preprints in each of the three fields of experimental high-energy physics, theoretical high-energy physics, and technological articles relevant for technology transfer. These have all been keyworded using the manual method.

The object of the test is to compare the automatic procedure with the manual method of keywording in use at DESY. We intend to find out whether the automatic procedure can be tuned to deliver keywords of a similar quality as DESY does with the manual method. Even if this turns out to be unrealistic, the expert system could be used as a Machine Aided Indexing (MAI) system to propose keywords to the indexer.

When the program has been tuned on these samples, it will be supplied with other samples of the same size. In comparison with the manually extracted keywords, the keywords extracted automatically should be as similar as possible in quantity (number per abstract) and in character.

70 000 words have been used as the "learning text" so far. Among the last 4000 words processed, only 200 words were unknown to the existing dictionary and required an input from an operator.

250 rules have been defined. Any sentence, parsed through these rules, will end up with one or many possible noun phrases.

The thresholds are still being defined: they need to be regularly adjusted as more documents are processed.

CONCLUSION

We can draw three main conclusions from the analysis and tests done so far:
1) The necessity of automatically keywording High Energy Physics Grey Literature is obvious.
2) We are optimistic that we will be able to build a valid knowledge base of noun-phrases, using Sokrates.
3) It is not yet clear how difficult it will be to successfully map this base onto the HEP Index thesaurus.

In all cases (whether the mapping is robust or not), the idea is that whenever a new document (with an abstract) is entered into the system, the expert system quickly delivers a set of key terms. This output can be added to the database straight away or it can be mapped to the thesaurus to try to deliver an "assigned term" and finally, it can be checked by experts before being loaded.

If the CERN test is successful, it will be run on a large scale in order to progressively cover all HEP preprints not yet classified. In addition to the traditional keyword searching option, a new utility will need to be developed to offer searchers a simple way of browsing through the thesaurus.

REFERENCES

[1] http://weblib.cern.ch/Preprints
[2] http://www.desy.de/library/homepage.html
[3] http://www.slac.stanford.edu/spires/hep/
[4] http://www.iaea.or.at/programmes/inis/inis.htm
[5] http://www.iee.org.uk/publish/inspec
[6] http://aip.org/pubservs/pacs.html
[7] http://weblib.cern.ch/share/udctree/udctree.html
[8] Wright, L.W., Grossetta Nardini H.K., Aronson, A.R., Rindflesch , T.C (1999) " Hierarchical Concept Indexing of Full-Text Documents in the Unified Medical Language System Information Sources Map", Journal of the American Society for Information Science, vol 50 No 6, pp. 514-523
[9] Silvester, J.P. (1997) "Computer Supported Indexing: a History and Evaluation of NASA's MAI System", Encyclopedia of Library and Information Science, vol 61 Jan 1998, pp. 76-90
[10] Sokrates Learning System, Dr. I. Steinacker, Intelligent Text Analysis and Management, Jenewein Burotechnik Innsbruck Austria.