High Energy Physics Libraries Webzine

Issue 5 / November 2001

http://library.cern.ch/HEPLW/5/papers/3/

Experiences in Automatic Keywording of Particle Physics Literature

Arturo Montejo Ráez, David Dallman (*)

Abstract:

Attributing keywords can assist in the classification and retrieval of documents in the particle physics literature. As information services face a future with less available manpower and more and more documents being written, the possibility of keyword attribution being assisted by automatic classification software is explored. A project being carried out at CERN (the European Laboratory for Particle Physics) for the development and integration of automatic keywording is described.

Introduction

Bibliographic databases originally developed out of the traditional library card catalogues, which enabled users to access documents via various types of bibliographic information, such as title, author or report number. In addition these catalogues sometimes contained subject-related information, for example that provided by using the Universal Decimal Classification (UDC) scheme [McI].

Following the introduction of the first e-print archive in the particle physics (also known as high-energy physics) community in 1991 [Gin91], a collection of around 200,000 fulltext documents covering several fields of physics, mathematics and accelerator engineering is now freely available via the World Wide Web [SG96].

The catalogue (database) provides the interface between the user and the document being searched for. This functionality has been enhanced, but not yet changed in any of its essentials, by having the documents available in electronic form, making it possible to access them remotely.

Searching for information

There are two main uses of a bibliographic database. One is, when one is searching for a specific document which one already knows exists, wants to find out if it is available, and if so, get access to it. This is the so-called *referral approach*, somewhat like looking up a piece of information in an encyclopedia. One knows it is there, one just wants to see it.

The other main use is when one has a specific topic in mind, and wants to find documents which address this topic. It is only with this second type of use that we are concerned here. This implies a *subject-based approach* to a document collection.

While the referral approach is well supported by current database systems, the subject-based one is less well developed since it has to deal with knowledge content and heuristics. The attribution of metadata by a subject specialist is itself quite a complex process. Computing approaches to the same task are still far from providing viable alternatives to the subject expert. Nevertheless, tools might be able to be developed to make this task easier.

Adding value by subject indexing

Providing subject-correlated metadata to add value to database records is called subject indexing. Some articles may contain subject-oriented information supplied by the authors (though often only when the journal makes it a condition of acceptance!). For example, some journals use keywords, while others have adopted the Physics and Astronomy Classification Scheme (PACS) supported by the American Institute of Physics [oP]. However, these approaches are not used widely enough to be useful for global searching. Therefore, any globally useful data describing the subject content have instead to be supplied by the creators of the database used for the searching.

In libraries, indexes and catalogues have traditionally been used to label documents in such a way that a user could search using this added metadata in order to find documents related to his requirements. Using a subject tree, that is a hierarchy of subject categories, he could choose to follow a path until he arrived at a branch with a limited set of documents, at which point he could then start a deeper search.

Another approach to subject indexing is to augment the database record with keywords or keyword phrases. These metadata are added to the database records in order to improve the quality of the search results, because searching the titles alone is likely to result in many of the relevant documents being missed. If abstracts can also be searched, more of the relevant documents will probably be found (higher recall) but at the expense of a reduced precision because even more non-relevant documents are likely to appear. This situation becomes even more extreme if a search in the fulltexts of the documents is possible. All these search possibilities do in fact exist on the CERN Document Server [DG96] for most particle physics documents since 1994. With the keywords one tries to synthesize the content of the document in relatively few terms (typically ten to twenty), in order to maximise both the precision and the recall, which are two negatively-correlated concepts when data coming only from the documents themselves are searched (whether it be titles, abstracts or fulltexts).

There are two distinct ways of carrying out the keywording process: terms can either be chosen from a predefined thesaurus or they can be selected at will by the indexer. 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 exact indexed terms need not even necessarily appear in the text at all, which can give this method an advantage over any strategy based only on the text of the document. Examples of fixed thesauri in the physics field are those used by INIS [INI] (International Nuclear Information System, Vienna), INSPEC [INS] (Physics, Computing and Electrical Engineering Abstracts, UK), and, in the particle physics field, at the Deutsches Elektronen-SYnchrotron laboratory (DESY) in Hamburg [DES96].

DESY Keywords

At DESY, the documentation group developed HEPI (the High Energy Physics Index) from 1963 onwards. In this scheme all documents in the particle physics field are indexed by subject specialists. The thesaurus used contains approximately 2500 terms and has in general been updated every one to two years. The DESY keyword terms are included with each record in the Stanford Linear Accelerator Center (SLAC) HEP database [SL75].

Types of keyword

There are three types of keyword or keyword phrase in the DESY thesaurus:

1. main keyword
2. descriptive keyword
3. non-keyword

These keywords can be used singly or in pairs. For example: `field theory' (single keyword phrase) or `field theory, nonabelian' (coupled pair of keyword phrases). While the first term is generally a main keyword, the second term may be a main keyword, a descriptive keyword or a non-keyword. Non-keywords that are used frequently are standardized however.

For the purposes of the automatic indexing project described here, only main DESY keywords are considered. The following is an extract from the DESY thesaurus. Keywords preceded by a blank are main keywords, those labelled with ``*'' are descriptive (secondary) keywords and those with ``-'' are non-keywords.

*coherent interaction
coherent state (for quantum mechanical states)
cohomology
*coil
-coincidence ('fast logic' or 'trigger' or 'associated production')
-Coleman-Glashow formula (baryon, mass difference)
-Coleman-Weinberg instability (symmetry breaking)
*collective (used only in connection with accelerators)
*collective phenomena ('field theory, collective phenomena' or 'nuclear physics, collective phenomena' or 'nuclear matter, collective phenomena')
-collider ('storage ring' or 'linear collider')
colliding beam detector (use only in instrumental papers)
*colliding beams (for accelerator use 'storage ring' or 'linear collider')
color (for colored partons)
colored particle
communications

Automatic Indexing

Automatic indexing based on word frequency can be traced back to the 1950s and the work of Luhn [Luh57] and Baxendale [Bax58]. Several studies have appeared since then, introducing more sophisticated techniques such as conflation algorithms and weight normalization. Many approaches have been used to try to provide an automatic solution to the indexing problem. But since detailed

subject knowledge is required in this process, there is no satisfactory fully-automatic indexing system available at present. At best, automatic systems have to be supervised by experts in order to control and adjust the results obtained.

Approaches

The electronic availability of large collections of fulltext documents has signalled the beginning of a new era in information retrieval. Much research is being done around natural language processing, to which the early work of Salton [Sal69] provides a good introduction. There has been a symbiosis between the fields of information extraction, text analysis and computational linguistics with the aim of providing solutions in this new environment. However, improved tools need to be used for the text processing, and automatic keywording still has some way to go until a successful fully automatic solution is arrived at. Automatic keywording can be considered to be a branch of automatic summarization, which aims at the generation of abstracts from fulltext documents. Many relevant algorithms have been developed in this approach, from classic conflation algorithms to reduce the representation of a document to its essentials (see [RW93]), to those which treat the document as a whole, identifying discourse trees [Mar97] or conceptual phrases [Cul83]. Examples of such systems are: BIOSIS [VS87], MeSH [Nat93], the NASA MAI System [GH98] and the indexer used by the European Commission for multilingual purposes [Ste01].

At CERN (the European Laboratory for Particle Physics) near Geneva there was a recent attempt to incorporate software for automatic indexing [Meu99]. The approach used first had a learning phase in which a sample of documents was used to train the system, which worked by constructing a set of the longest noun-type phrases from the abstracts of the documents. The system worked automatically, and did not require experts to read the articles to be keyworded. Once the training period was over it was completely autonomous, that is, it was not intended to be merely a tool to aid indexers.

After some tests, this project was abandoned at the end of 2000, mainly because of the difficulty of integrating existing non-modular software into the database system of the CERN Library. Instead, a new project, whose purpose was to develop an indexing tool which would supply DESY keywords directly, was launched. This project is called *HEPindexer*.

HEPindexer

The CERN Scientific Information Service harvests over a hundred documents each day from electronic sources all around the world. Owing to the growth in the amount of literature, a new approach to keywording has been developed. Accepting the fact that fully automatic indexers are still far from providing a complete solution, a computer-based help tool for indexing was developed as a first step towards easing the work of human indexers.

Thus *HEPindexer* proposes a preliminary solution which can open the way for further research into automatic indexing tools in the area of particle physics. So far a first step has been achieved, namely the generation of *main DESY keywords*. These keywords are generated following a statistical approach [vR75].

The algorithm used needs a set of data which must be generated in a *training* process beforehand. HEPindexer was supplied with a training sample of about 2400 particle physics documents together with the DESY keywords that had already been assigned to these documents. It uses the fulltext of these documents. After training, a new document can be submitted to the system and receive as

output the list of automatically proposed DESY keywords. A sample of 1200 additional documents was used to evaluate the results of the algorithm. The results were close to 60% in both precision and recall. This means that an average of about 60% of the keywords proposed by HEPindexer are also contained in the list proposed by DESY, and that about 60% of the keywords proposed by DESY are among those proposed by HEPindexer.

HEPindexer has now been integrated into the CERN Document Server (CDS)[DG96]. By retrieving any document in the database for which the fulltext is stored on the CDS, one can then run HEPindexer and obtain the list of proposed main DESY keywords. To do this, choose "Fulltext" and then click on the text "10 main keywords". By clicking on "Corresponding record at SLAC" one can directly see the keywords which were actually supplied by DESY and make a comparison. If the link to the SLAC record is not present, one can instead go to the SLAC database and retrieve the same record. Note that for recent records the DESY keywords may not yet be available.

Improvements in HEPindexer are still being made, as the project is only in its initial phase. Secondary keywords and more refined algorithms (using linguistic resources) are being studied in order to enhance the performance of the system. At present, each main keyword proposed simply has a link to the list of secondary keywords which can be used in association with that main keyword. The indexer can then choose from this list the secondary keywords which are relevant.

HEPindexer is conceived as a tool to help an indexer to find the best keywords for particle physics documents. It is still not possible to replace the human task by a fully automatic tool, but it can make it easier and faster.

Acknowledgements

This project was carried out in the ETT division of CERN by Arturo Montejo Ráez, under the supervision of Jean-Yves Le Meur and Jens Vigen.

Bibliography

Bax58 P. Baxendale. Machine-made index for technical literature -- an experiment. *IBM Journal, pages 354-361, October 1958.*

Cul83 Christopher Culy. An extension of phrase structure rules and its application to natural language. Master's thesis, Stanford University, 1983.

DES96 DESY. The High Energy Physics Index keywords. http://www-library.desy.de/schlagw2.html

DG96 CERN. DH group, ETT division. The CERN Document Server. http://cds.cern.ch/

GH98 Oak Ridge, Gail Hodge. CENDI agency indexing system descriptions: A baseline report. Technical report, CENDI, 1998. http://www.dtic.mil/cendi/publications/98-2index.html

Gin91 Paul Ginsparg.
arXiv.org e-print archive.
http://www.arxiv.org/

INI INIS thesaurus.
http://www.iaea.or.at/worldatom/publications/inis/inis.html

INS INSPEC thesaurus.
http://www.iee.org.uk/publish/inspec/

Luh57 H. Luhn.
A statistical approach to mechanized encoding and searching of literary information.
IBM Journal of Research and Development 1 (4) 309-317, 1957.

Mar97 Daniel Marcu.
Discourse trees are good indicators of importance in text.
Technical report, Information Science Institute, University of Southern California, 1997.

McI I.C McIlwaine.
The Universal Decimal Classification - a guide to its use.
Technical report, UDC Consortium.

Meu99 D. Dallman; J. Y. Le Meur.
Automatic keywording of high energy physics.
In *4th International Conference on Grey Literature : New Frontiers in Grey Literature*, Washington, DC, USA, Oct. 1999.

Nat93 National Library of Medicine, Bethesda, US.
Medical Subject Headings (MeSH), 1993.

oP American Institute of Physics.
Physics and Astronomy Classification Scheme.
http://publish.aps.org/PACS/

RW93 A. M. Robertson and P. Willett.
Evaluation of techniques for the conflation of modern and seventeenth century english spelling.
In Tony McEnery and Chris Paice, editors, *Proceedings of the BCS 14th Information Retrieval Colloquium*, Workshops in Computing, pages 155-168, London, April 13-14 1993. Springer Verlag.

Sal69 Gerard Salton.
Automatic text analysis.
Technical Report TR69-36, Cornell University, Computer Science Department, June 1969.

SG96 CERN. SI group, ETT division.
CERN Scientific Information Service.
http://library.cern.ch/

SL75 Stanford Linear Accelerator Center. SLAC Library.
SLAC SPIRES HEP database.
http://www.slac.stanford.edu/spires/hep/

Ste01 Ralf Steinberger.
Cross-lingual keyword assignment.
In L. Alfonso Ureña López, editor, *Proceedings of the XVII Conference of the Spanish Society for Natural Language Processing (SEPLN 2001)*, pages 273-280, Jain (Spain), September 2001.

vR75 C. J. van Rijsbergen.
Information Retrieval. London: Butterworths, 1975.
http://www.dcs.gla.ac.uk/Keith/Preface.html

VS87 Natasha Vieduts-Stokolo.
Concept recognition in an automatic text-processing system for the life sciences, 1987.

Author Details

Arturo Montejo Ráez, David Dallman
Scientific Information Service
CERN - ETT Division
Geneva (Switzerland)
***Email**: David.Dallman@cern.ch*
***Email**: Arturo.Montejo.Raez@cern.ch*
***URL**: http://library.cern.ch/*

For citation purposes:
Arturo Montejo Ráez, David Dallman, "Experiences in Automatic Keywording of Particle Physics Literature", High Energy Physics Libraries Webzine, issue 5, November 2001
URL: <http://library.cern.ch/HEPLW/5/papers/3/>