



*High Energy Physics Libraries
Webzine*

[Home](#)

[Editorial Board](#)

[Contents](#)

[Issue 10](#)

HEP Libraries Webzine

Issue 10 / December 2004

Why Keywording Matters

Arturo Montejo Ráez, Ralf Steinberger(*)

Abstract

Most information retrieval systems nowadays use full-text searching because algorithms are fast and very good results can be achieved. However, the use of full text indexes has its limitations, especially in the multilingual context, and it is not a solution for further information access requirements such as the need for efficient document navigation, categorisation, abstracting and other means by which the document contents can be quickly understood and compared with others. We show that automatic indexing with controlled vocabulary keywords (descriptors) complements full-text indexing because it allows cross-lingual information access. Furthermore, controlled vocabulary indexing produces document representations that are useful for human users, for the Semantic Web, and for other application areas that require the linking and comparison of documents with each other. Due to its obvious usefulness, controlled vocabulary indexing has received increasing attention over the last few years. Aiming at a better understanding of the state-of-the-art in this field, we discuss the various approaches to automatic keywording and propose a taxonomy for their classification.

Introduction

We launch our web browser and, after clicking on a bookmark, a one-field form appears embedded in the page. Once a few words are typed inside the text field, we click on the 'submit' button expecting the answer to our question. A few seconds later the browser shows a page containing a list of items: those the system considers most suitable to answer

our needs. The discrimination of results becomes a non-trivial operation due to the large number of entries returned. Sometimes we can get rid of some of them at a glance: the title or the text provided along with the item is enough to know we are not interested, but sometimes we have to click and check the real document to see whether it is the information we want, or not.

Many of us will recognize the sequence of steps performed above. We were looking for information using a *full-text* search engine. This operational mode in information searching and retrieval has populated almost every digital system which stores information. We can find forms like the one described when:

- Searching for web pages: as in the example above. Search engines are amongst the biggest aggregators of information nowadays.
- Searching for files: most operating systems come with tools supporting this feature. Thus, we can search for files containing certain words. Some systems even allow the possibility of using regular expressions, that is, as a more advanced form of the useful **grep** command on UNIX systems. Also commonly-used electronic mail clients let the user look for a message containing a particular word in a collection of email messages.
- Searching for books: now libraries offer their catalogues on-line and, in the case of electronic libraries, they can search for query words amongst the full text of documents stored.
- Searching for reports: some administrative tools integrate inverted files into their structure to make searching faster.
- and more...

Though the usefulness of full text search engines has been widely proven and, therefore, accepted, they are still not good enough in some cases and totally inappropriate in others. The first kind of less-successful cases are those where the collection contains a huge range of subjects and documents: for example, the World Wide Web. Old approaches using purely full-text-based engines were abandoned, since the quality of results provided was declining with the growth of the collection. Therefore, new techniques arose with the aim of filtering and re-qualifying the rank (the Page Rank algorithm is one of the most successful examples [1]). They index every word in a page so they can perform full-text searches later. The problem with this approach is that language is complex, ambiguous and rich in variation, so the quality of the results is still not as good as we would like. But this technique of indexing is solving the big problem of searching for information on the web. It is an implementable solution in very general contexts.

The second field where full text-search techniques do not do so well is when textual information is not available. There are still some kinds of collections which are not suitable (yet) for this genre of engines. We refer here to pieces of information like images, sounds, etc. The current solution is to provide, beforehand, textual information related to every item (that is, enrich the data with text) so that later we can search using this related text as an access point. Many techniques have been developed in order to automate such a process by pattern recognition, clustering and so on.

Subject keys in traditional information systems

Imagine you had to organize your personal library, what sort of ideas do you think you would try in order to achieve well organized shelves? Maybe one of your first ideas would be to group books by theme, then to label them and put their details in a kind of index. Later on you might find you have so many books, it would be better to arrange them by size (large repositories do so). Whatever method you used, in the end you would have to *index* them in one way or another. Now the question could be: which indexes should I use? It is not an easy task to define them because several considerations must be taken into account. Vickery already emphasizes this reality [2]:

The problem of subject representation is therefore far less straightforward than other aspects of document description.

In the beginning, the use of keywords for information storage and retrieval was due to two major needs: the need for classification and the need for retrieval. The former need had a double benefit: first, it let librarians organize physical volumes into logical clusters; second, the possibility to search within a defined cluster was regarded as a way to speed up the searching for information (as pointed out by the so-called 'cluster hypothesis' of Rijsbergen [3]).

Hence, two major goals of indexing are to:

1. Select records in a file that deal with a specific topic
2. Group in proximity in a file records on similar subjects

Alphabetical terminologies and classification structures (known as 'thesauri') were thought of as tools to improve the two main *measures* in information retrieval: *precision* and *recall*. These refer to the quality of retrieved documents when compared to the search query. 'Precision' is the number of relevant documents retrieved over the total number of documents retrieved. 'Recall' is the number of relevant documents retrieved over the total number of relevant documents in the collection. These two measures show the problem of an antagonistic relationship: if we try to improve one of them, the other will decay. For example, if we retrieve the whole collection in answer to a given query, our recall will be 100%, but our precision will be so low that the result will be unusable. The challenge resides, then, in finding a method which shows a good performance for both measures.

In earlier times, techniques were used to improve these two values for a defined retrieval system; i.e. the implementation of these techniques was oriented to the purpose and content of the retrieval system. The techniques traditionally used rely on setting relationships between words in a controlled vocabulary. Using those relations on a given query we can improve recall (by expanding to related terms) or precision (by narrowing with less generic terms). These are the reasons for the use of thesauri.

Thesauri

There are several definitions for the word 'thesaurus'. In an old work of Vickery [2] we

find a definition for thesaurus which summarizes in a few words the rationale associated with it:

"The thesaurus is a display of the terms in a retrieval language showing semantic relations between them."

Here, Vickery shows, on the one hand, the main purpose of a thesaurus: it defines a retrieval language, whatever the retrieval method might be. On the other hand, he does not define the kind of relationships between entries (synonyms, broader terms...), specifying only that a set of semantic relations is defined. We will see that this brief definition fits perfectly with any type of existing thesaurus.

One of the earliest thesauri (and maybe the most famous one) is *Roget's Thesaurus* [4]. Dr. Peter Mark Roget's main idea behind this compilation was to create a system which would offer words to express a given meaning, while conversely traditional dictionaries offer meanings for a given word. This would help writers to express their concepts in the most suitable form. Such users had the thesaurus as a reference book for writing of texts. Thus, it was mostly intended to be useful in the document creation phase.

The power of reducing a language to its basic concepts has become more and more useful, especially since the "semantic network" has arisen in electronic form. WordNet [5] is an on-line reference system" (their authors state). English nouns, verbs, adverbs and adjectives are organized into *synonym sets* (also called *synsets*), each representing one underlying lexical concept. Nowadays we can assume that almost every thesaurus (specialized or not) is available in electronic form.

Thesaurus descriptors are normally unambiguous because they are clearly defined, whereas full text indexing does not provide any differentiation for words such as 'plant' in 'power plant' versus 'green plant'.

There is even a multilingual thesaurus based on WordNet called EuroWordNet [6], which, using English as central node, maps synsets between different European languages. This work represents a milestone in multilingual information retrieval.

Both WordNet and Roget's Thesaurus are general reference sources, i.e. they don't focus on specialized terminologies. But the areas where thesauri become useful tools are in specialized domains (Law, Medicine, Material Science, Physics, Astronomy...). One example is the INSPEC thesaurus [7], focused on technical literature; or the ASIS thesaurus, specialized in Information Science [8]. NASA, the European Union, and other organizations produce their own specialized thesauri (like the multilingual EUROVOC thesaurus [9]).

Each thesaurus has its own organization, according to the purpose it needs to accomplish. But we can summarize any of them by the following components:

Terms

the set of items in the thesaurus. They are usually referred to as descriptors, index terms, keywords, key phrases, topics, concepts or themes. We will use "keyword" to name them.

Meanings

the set of subsets of the set of terms. Each subset contains a group of terms which are interrelated by the *synonym* relationship (i.e. words with the same meaning). This relationship is important because resulting subsets are elements in other relations.

Relationships

this is a set of relations keyword to keyword, keyword to meaning, meaning to keyword and meaning to meaning.

There are two relations which are commonly used among existing thesauri:

- **Hyponymy.** This is a relationship between meanings. We say that *x is a hyponymy of y* if *x is a kind of y*. This relation is reflexive, anti-symmetric and transitive, therefore it establishes a *partial order* between meanings. The symmetric relation is called *specialization* and also defines a partial order over the set of descriptors.
- **Meronymy.** This can be split into three different (but closer) relationships:
 1. *x is part of y*, e.g. *branch* is part of *tree*
 2. *x is a member of y*, e.g. *citizen* is member of *society*
 3. *x is constituent material of y*, e.g. *iron* is constituent material of *knife*

Of course, depending on the purpose of the thesaurus, some of these relations may be ignored. Also new relations could occur. WordNet, for example, includes all of the given relations. INSPEC and Eurovoc thesauri condense meronym relations into the "related" relationship (see [8], *RT* means "related terms"). Synonymy is implemented by the application of the "USE" statement.

```
[...]
penalty
NT1  alternative sentence
NT1  carrying out of sentence
      NT2  barring of penalties by limitation
      NT2  reduction of sentence
      RT  repentance
      RT  terrorism      (0431)
      NT2  release on licence
      NT2  suspension of sentence
NT1  conditional discharge
NT1  confiscation of property
      RT  seizure of goods      (1221)
NT1  criminal record
NT1  death penalty
NT1  deprivation of rights
      RT  civil rights      (1236)
[...]
```

Figure 1: Excerpt from Eurovoc thesaurus

Usually in specialized thesauri either the synonymy is neglected or a preferred word representing the meaning is given, since the purpose is to provide a list of controlled terms (and that "control" refers to the use of just one word for a given meaning).

Nevertheless, most of them include synonymy in one way or another.

There are, however, some special cases of thesauri where there is more than just terms and relations. In some cases the thesaurus is a complex reference of specific relations, with specially defined rules to build a document's keywords. This is the case of the DESY thesaurus [10], specializing in high energy physics literature. With the entries given we can construct particle combinations, reaction equations and energy declarations among other examples.

These facts bring us to the conclusion of Vickery that there is a tight relationship between the thesaurus and its domain of retrieval.

Applications of keywords

The construction of hand-crafted thesauri for use in computer applications dates back to the early 1950s with the work of H. P. Luhn, who developed a thesaurus for the indexing of scientific literature at IBM. The number of thesauri and systems is now growing steadily because manually or automatically keyworded documents have many advantages and offer additional applications over simple documents not linked to thesauri. Depending on whether people or machines make use of keywords assigned to documents, we distinguish the following uses:

Human manipulation of keywords

Human users mainly use keywords for *browsing* and *searching* of document collections.

Browsing

Keywords are used to facilitate the browsing of document collections, either as part of a whole collection or the small subset returned by a search operation. Examples of how keywords can aid browsing:

- **Use of keywords as a document summary.** Thesaurus descriptors are usually a small list of carefully chosen terms that represent the document contents particularly well. Depending on the thesaurus, they are of a summarising, conceptual nature. They often do not occur explicitly in text so that they are of a completely different nature from full text indexes. Descriptors function as a kind of abstract summary and give users a quick and rough idea of the document contents. This helps the users to quickly sieve out the most important or relevant documents from a large collection. The use of keywords as a means for automatic summarization is an interesting application already in practice in many digital libraries and on-line catalogues.

For high energy documents this can speed up the search process in specialized collections which grow by hundreds of documents every week [11].

If the thesaurus is multilingual, this summarising function also works across languages, i.e. a user will see a list of keywords (a summary) in their own language

of a document written in another language [12].

- **Use of keywords for document navigation.** If the database containing the full texts and their keywords offers hyperlinks based on the keywords, it is possible to navigate through the document collection by starting with one document and searching for similar documents by clicking on one or more of the keywords to see other documents indexed with the same descriptors.
- **Classification of documents.** In some search engines the results of a search are classified "on the fly" into categories so that the browsing of documents is easier and more self-organized. The user can distinguish faster between interesting documents and irrelevant ones. Although many of these systems use words from documents to label automatically generated categories, others select them from a controlled vocabulary. An example for such a system is GRACE [13].

However, the use of keywords for classification goes beyond the pure retrieval domain. The *freedesktop.org* project [14] promotes the use of keywords for the arrangement of icons (representing application launchers) in the main menu of the desktop where applications are internally attached to a list of categories. It means that there is no predefined taxonomy to which program launchers are classified. Instead, programs are labelled with keywords from which menus are created in the graphical interface of an operating system (like the Gnome [15] desktop available for Linux and other operating systems).

- **Identifying the common subject of cited documents.** An interesting feature that will be tested under the HEPindexer project is the use of keywords to help users in navigating through document references, allowing them to recognize the subject that is shared by the reference and the document the reference belongs to.

Searching

Keywords are helpful during the search phase. For example:

- **For query expansion.** Some authors, like Vassilevskaya [16] among others, propose the use of controlled vocabulary for query expansion. In the query formulating process, the query is passed to the automatic assignment tool and some thesaurus keywords are suggested to the user. These can then either be chosen instead of, or in addition to, the query.
- **For cross-lingual searching.** When the thesaurus used for indexing is multilingual, users can be given the option to use thesaurus descriptors as search terms. The search can then be carried out using search terms in another language to achieve cross-lingual document search and retrieval.
- **Descriptors provided by a thesaurus** have a relevant advantage over single terms selected for automatic full-text indexing. Concepts which can be expressed in several synonymous ways, such as "Chemotherapy" and "Drug Therapy", are conflated to only one form ("Drug Therapy"), while phrases can be treated as single concepts ("stage IIIB breast cancer", for example). This is one of the conclusions found by using the MetaMap Indexer for medical documents [17].

- When searching in highly specialized domains, keywords can be used as a directory or subject tree for the user who is not able to make his/her information needs explicit as a set of query terms. If documents in the collection have been labelled with keywords and the structure of the thesaurus is hierarchical, the user can drill down through the categories narrowing the search space. This could be the electronic equivalent of browsing the Universal Decimal Code (UDC) when we enter a library for the first time. Such a tree could profit from semantic relations between keywords, hence we may find related topics, general topics, and further relationships.

Using keywords as a document representation for machine usage

The fact that keywords were traditionally developed for human readers does not necessarily mean that they can only be used by people. Several powerful applications have shown that descriptors can well be used to represent document contents for a number of automatic procedures:

- **Using descriptors from a hierarchically organised thesaurus allows searching by subject field.** For instance, searching for "RADIOACTIVE MATERIALS" can automatically be extended to all individual instances of radioactive elements, such as 'uranium', 'plutonium', etc. This form of query expansion is being used for some domains like high energy physics [16].
- **Multilingual document similarity calculation.** As the list of thesaurus descriptors of a text is a semantic representation of this text, texts can be compared with each other via these descriptor lists. The idea is that the higher the number of content descriptors two documents have in common, the more similar are the documents. For multilingual thesauri like Eurovoc [9], which currently exists in twenty-one language versions with one-to-one translations for each descriptor, the document similarity calculation is even possible for texts written in different languages. Steinberger, Pouliquen and Hagman have shown [18] that a translation of a given document can quite reliably be identified in a multilingual document collection because it is correctly identified as the most similar document to the original.
- **Multilingual clustering and multilingual document maps.** The same cross-lingual document similarity measure can be used as input to further multilingual applications. These include multilingual clustering and classification of documents, as well as the visualisation of multilingual document collections in single document maps [19].
- **The Semantic Web.** Our opinion is that all these technologies will play a key role in the development of the Semantic Web [20]. Considering that the whole structure of the Semantic Web depends on RDF (*Resource Description Framework*) and that there are already some projects to use thesauri like a schema (ontology) defining the terms used to represent the RDF version of WordNet, we can conclude that this is a promising area of research. A web of documents can be related via their associations between keywords.

- **The Semantic Grid.** Not only documents can be linked using keywords, but any type of service (for example, any web service) can be attached with keywords which could be automatically assigned using the description of the service as a basis. The Semantic Grid [21] objective can be reached faster using subject enhancement, i.e. keywording.

Therefore, we could imagine a scenario where we want to look for a service, e.g. a database of iron manufacturers. We get the keywords of the service which may have been generated from the content of top web pages in the portal of this service (the pages which let us access the database via web forms or any other web based interaction). These keywords show us that there is another database which offers iron toys, since the thesaurus splits the keyword *iron manufacturer* into the subtopics *iron-made toys manufacturer*, *naval manufacturer*, etc. Thanks to the semantic network created from keyword relationships we are able to find the provider we need.

Automatic key word assignment tools: a taxonomy

F. W. Lancaster [22] gives us the following definition for *indexing*:

"The main purpose of indexing and abstracting is to construct representations of published items in a form suitable for inclusion in some type of database."

This is a very general description, but it still summarizes the goal of indexing: provide an abstraction of document contents for better storage and retrieval (which is the goal of any database). We find several types of indexing. For example, web search engines (like Google, Altavista and others) generally full-text-index web pages automatically, but for some specialized and popular subject areas, they ask teams of professional indexers to carry out careful manual indexing.

We distinguish two main types of indexing [22]:

- **Indexing by assignment.** Most human indexing is assignment indexing, involving the representation of subject matter by means of terms selected from some form of controlled vocabulary. Due to the intellectual work involved, this manual task is very labour-intensive and therefore expensive. Fortunately, new automatic solutions are emerging with reasonable performances.
- **Indexing by extraction.** Words or phrases appearing verbatim in a text are extracted and used to represent the contents of the text as a whole. Keyword extraction is thus less abstract and more limited than assignment indexing.

The taxonomy we propose focuses on those systems that do *assignment* of keywords instead of extraction.

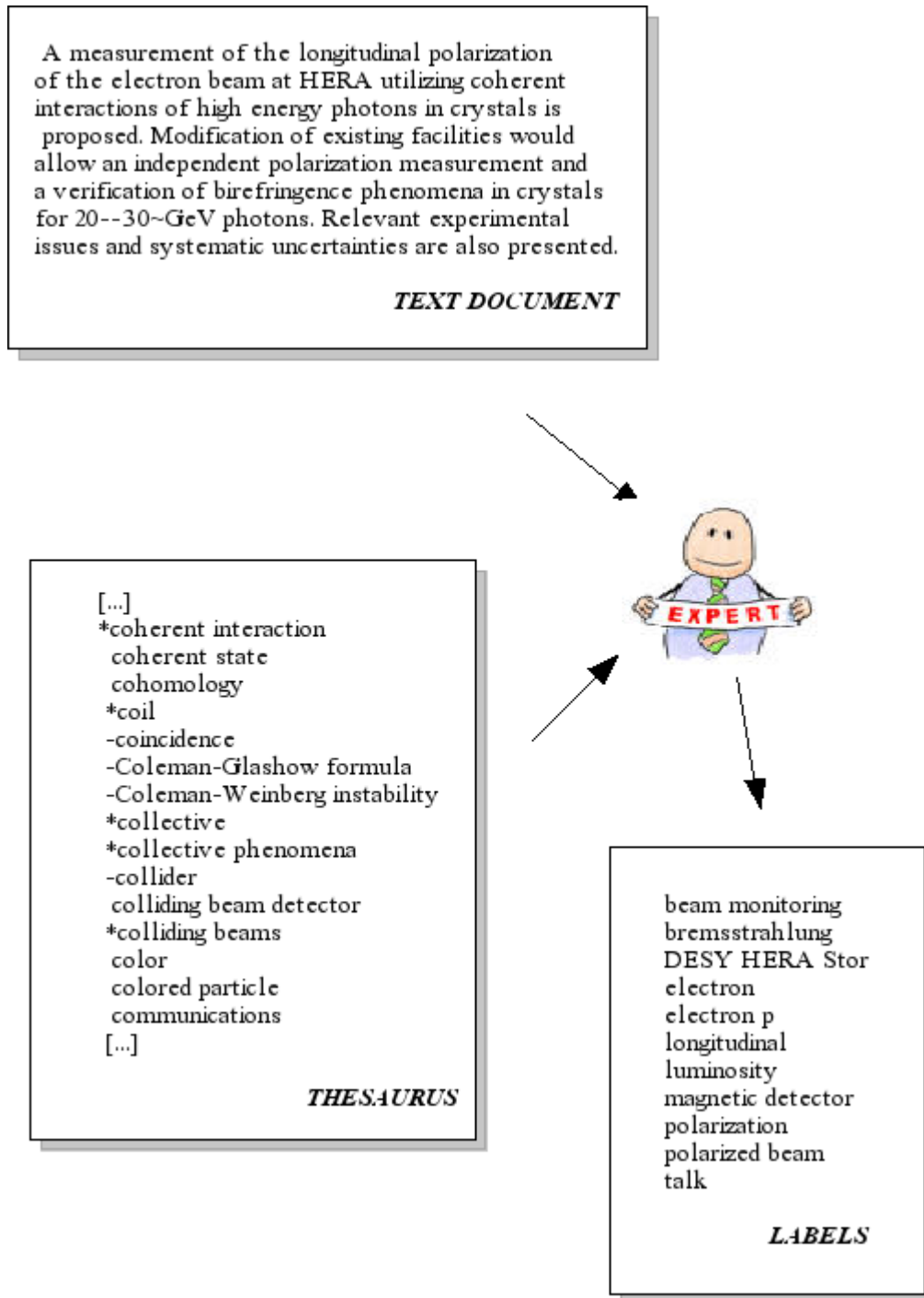


Figure 2: Indexing by assignment

Figure 2 provides a graphical view of this process. The indexer reads the document and selects keywords from a thesaurus or a controlled vocabulary.

Although there has been extensive work done on automatic thesauri generation, less work has been done on automatic descriptor assignment. Although research is advancing slowly in this area, it benefits from development in other IR (information retrieval) areas. We mention here some of the systems developed for automatic assignment, along with a taxonomy proposal for this kind of tool.

We can classify automatic keyword assignment systems (AKWAs) into two main categories, depending on their degree of automation:

- **Machine Aided Indexing (MAI):** those systems *supporting* indexers in their manual task of finding good keywords for documents, like NASA MAI System [23] or BIOSIS [24].
- **Fully Automatic Indexing (FAI):** those systems intended for a *fully automatic* keywording assignment process without any human intervention. For FAI tools a document is used as input and the system automatically produces as output a list of keywords.

The following is a taxonomy that summarizes the various AKWA approaches developed so far:

- **Indexing by analogy to similar documents.** This approach can be used if there is already a collection of pre-indexed documents with which a new document can be compared. The basic idea is that the document, for which indexing is to be carried out, should be indexed with the keywords of the most similar documents in the collection. For this purpose, an existing document retrieval engine will identify the most similar documents in the collection. In the following step, the most frequent keywords of the retrieved documents will be merged and assigned to the new document. The process of retrieving similar documents in the collection can be based on a lexical vector space model, on the references the documents have in common, on formatting features [25], or on any other similarity measure.

Advantages of this relatively simple approach are its relatively easy implementation and the fact that, depending on the kind of merge function used, training may not be necessary. A disadvantage is that the merging function might be difficult to implement.

An example for this category is the system developed by Ezhela et al. [26], which is fully based on the references linked to a publication. Even with a quite simple approach, the results were acceptable, due to the high specialization of the subject domain (High Energy Physics). The merging strategy used was a pure intersection operation.

- **Indexing by classification.** This approach also requires the existence of previously indexed documents. The idea is that each descriptor is treated like a class and there are as many classes as there are descriptors in the thesaurus. A document indexed with five descriptors is thus a document multiply classified into five classes. Each class is represented by all those documents that have been indexed with the corresponding descriptor. During the assignment phase, a new document will be classified into the most appropriate classes in order to identify the most appropriate descriptors. The quality of the assignment depends on the performance of the classification algorithm used.

Most current systems fall under this category. One of them is the first version of the HEPindexer system [27], which uses a vector space model fully dedicated to the automatic assignment of descriptors.

Finally, the last criterion for classifying AKWAs is based on *training needs*, i.e. on the amount of effort required to develop the system:

- **Machine learning systems** always require automatic training using a model collection of documents that have previously been indexed. The systems can vary quite a lot depending on the type of document representation that is given to the text. This can range from a simple word frequency list for each document to a multi-faceted collection of document features. Typically, documents are represented by either all their words, or by all their lemmas, or by all their nouns and noun phrases.

An example of this kind of approach is the European Commission's EUROVOC indexing tool [28], which represents each class by a combination of the most frequent lemmatized keywords of all of its documents, uses a large number of parameters and calculates various types of vector space similarities between the new document and all of its classes in order to determine the most appropriate descriptor. A more standard approach to Machine Learning was taken by Hulth [29], but her work is limited to keyword extraction rather than assignment.

A general, positive feature of these systems is that they can rather easily be adapted to new thesauri and languages as they automatically optimise the selection of features from the feature set provided as input. Language-independent systems typically work with little linguistic input (e.g. only with a text's frequency list of word tokens). Better performance can be achieved using more linguistic processing such as lemmatization and parsing or using linguistic resources such as synonym lists and other dictionaries.

- **Rule based systems** typically make an intensive use of linguistic resources and use language- and/or domain-dependent rules which are normally developed manually. The major problems with this approach are its development cost and the fact that the systems cannot easily be adapted to new domains and languages. The work done by Vassilevskaya [16] fits under this category. She proposed a system specialized on high energy physics, based on five types of rules with hundreds of rules introduced manually. A more extreme example of this labour-intensive approach is that of Hlava and Hainebach [30], who produced over 40,000 hand-crafted rules for English alone. The conditions were of various types, including conditions regarding the presence of text strings, the usage of synonym lists, vicinity operators and even the recognition and exploitation of legal references in texts.

Conclusions

We have shown that manual or automatic indexing of document collections with controlled vocabulary thesaurus descriptors is complementary to full-text indexing and that it provides both human users and machines with the means to analyse, navigate and access the contents of document collections in a way full-text indexing would not permit. Indeed, instead of being replaced by full-text searches in electronic libraries, a growing number of automatic keyword assignment systems are being developed that use a range of very different approaches.

In this paper, we have given an introduction to automatic keyword assignment,

distinguishing it from keyword extraction, and proposing a classification of approaches, referring to sample implementations for each approach. This presentation will hopefully help researchers in the area to better understand and classify emerging approaches. We have also summarized some of the powerful applications that this kind of tool is offering in the field of information retrieval, and we have structured them into comprehensive categories and have shown real examples and working solutions for some of them.

As the number of different systems for automatic keyword assignment has been increasing over recent years, it was our aim to give some order to the state of the art in this interesting and promising field of research.

References

- [1] L. Page, S. Brin, R. Motwani, and T. Winograd. "The pagerank citation ranking: Bringing order to the web." *Technical report*, Computer Science Department, Stanford University, 1998.
- [2] B. C. Vickery. *Information Systems*. London: Butterworth, 1973.
- [3] C. J. van Rijsbergen. *Information Retrieval*. London: Butterworths, 1975.
URL: <<http://www.dcs.gla.ac.uk/Keith/Preface.html>>
- [4] S. M. Lloyd, editor. *Roget's Thesaurus*. Longman, 1982.
- [5] George A. Miller, Richard Beckwith, Christian Fellbaum, Derek Gross and Katherine Miller. *Introduction to WordNet: An On-line Lexical Database*. Cognitive Science Laboratory, Princeton University, August 1993.
- [6] P. Vossen. "Eurowordnet: a multilingual database for information retrieval" in: *Proceedings of the DELOS Workshop on Cross-language information retrieval*, Zurich, Mar 5-7, 1997.
- [7] *Inspec thesaurus*. Homepage
URL: <<http://www.iee.org.uk/publish/inspec/>>
- [8] *ASIS thesaurus of information science*. Homepage
URL: <<http://www.asis.org/Publications/Thesaurus/isframe.htm>>
- [9] *Eurovoc thesaurus*. Homepage
URL: <<http://europa.eu.int/celex/eurovoc/>>
- [10] *DESY. The high energy physics index keywords*, 1996.
URL: <<http://www-library.desy.de/schlagw2.html>>
- [11] A. Montejo-Ráez and D. 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/>>.
- [12] R. Steinberger. "Cross-lingual keyword assignment." In: L. A. U. na López, editor,

Proceedings of the XVII Conference of the Spanish Society for Natural Language Processing (SEPLN'2001), pages 273-280, Jan (Spain), Sept. 2001.

[13] *The grace engine*. Homepage
URL: <<http://www.grace-ist.org>>

[14] *The freedesktop project*. Homepage
URL: <<http://freedesktop.org>>

[15] *The gnome website*. Homepage
URL: <<http://www.gnome.org>>

[16] L. A. Vassilevskaya. *An approach to automatic indexing of scientific publications in high energy physics for database spires-hep*. Master Thesis, September 2002.

[17] L. W. Wright, H. K. Grossetta Nardini, A. R. Aronson and T. C. Rindflesch. "Hierarchical concept indexing of full-text documents in the unified medical language system information sources map." *Journal of Education for Library and Information Science* 50(6) :514-523, 1999.

[18] Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. "Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC", in: A. Gelloukh, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference CIC Ling 2002*. Springer, LNCS 2276, 2002.

[19] Ralf Steinberger, Johan Hagman, and Stefan Scheer. "Using thesauri for automatic indexing and for the visualisation of multilingual document collections.", in: *Proceedings of the workshop on Ontologies and lexical knowledge bases (Ontolex, 2000)*, pages 130-141, Sozopol, Bulgaria, sept 2000.

[20] T. Berners-Lee, J. Hendler and O. Lassila. "The semantic Web." *Scientific American*, 284(5) (May 2001) :34-43.

[21] T. H. Fran Berman, Geoffrey Fox, editor. *Grid Computing: Making the Global Infrastructure a Reality*. Wiley, 2003.

[22] F. W. Lancaster. *Indexing and Abstracting in Theory and Practice*. London: Library Association Publishing, 1998.

[23] O. R. Gail Hodge. "Cendi agency indexing system descriptions: A baseline report." *Technical report*, CENDI, 1998.
URL: <<http://www.dtic.mil/cendi/publications/98-2index.html>>

[24] N. Vieduts-Stokolo. "Concept recognition in an automatic text-processing system for the life sciences". *Journal of the American Society for Information Science*, 38 (4): 269-287, 1987.

[25] A. Anjewierden and S. Kabel. "Automatic indexing of documents with ontologies." In: B. Krose, M. de Rijke, G. Screiber and M. van Someren, editors, *Proceedings of BNAIC 2001 (13th Belgian/Dutch Conference on Artificial Intelligence)*. Amsterdam,

Netherlands, 23-30, 2001.

[26] V.V. Ezhela, V.E. Bunakov, S.B. Lugovsky, V.S. Lugovsky and K.S. Lugovsky. "Discovery of the additional knowledge and their automatic indexing via citations (in Russian). " In: *third All-Russian conference, Digital Libraries; Advanced Methods and Technologies, Digital collections (RCDL'2001)*, Petrozavodsk, sept 11-13, 2001.

[27] A. Montejo-Ráez. "Toward conceptual indexing using automatic assignment of descriptors." In: Stephano Mizzaro and Carlo Tasso, editors, *Personalization Techniques in Electronic Publishing on the Web: Trends and Perspectives, proceedings of the AM' 2002 Workshop on Personalization Techniques in Electronic Publishing*, Malaga, Spain, May 2002.

[28] Bruno Pouliquen, Ralf Steinberger and Camelia Ignat. "Automatic annotation of multilingual text collections with a conceptual thesaurus." In: A. Todirascu, editor, *Proceedings of the workshop 'Ontologies and Information Extraction' at the Summer School 'The Semantic Web and Language Technology' (EUROLAN'2003)*, Bucharest, jul 28 - Aug 8 2003.

[29] A. Hulth. "Improved automatic keyword extraction given more linguistic knowledge." In *Proceedings of the Conference Empirical Methods in Natural Language Processing (EMNLP'2003)*, Sapporo, Japan, July 2003.

[30] Marjorie M.K. Hlava and Richard Hainebach. *Multilingual machine indexing*.
URL: <<http://joan.simmons.edu/~chen/nit/NIT'96/96-105-Hava.html>>

Authors Details

Arturo Montejo-Ráez

IT-UDS
European Organization for Nuclear Research
Geneva
Switzerland

Tel: +41 (0) 22 767 3833

Email: arturo.montejo.raez@cern.ch

URL: <http://cern.ch/amontejo>

Ralf Steinberger

European Commission
Joint Research Centre
T.P. 267, 21020 Ispra (VA)
Italy

Tel: +39 - 0332 78 6271

Email: ralf.steinberger@jrc.it

URL: www.jrc.it/langtech

For citation purposes:

A. Montejo-Ráez and R. Steinberger, "Why keywording matters". *High Energy Physics Libraries Webzine*, Issue 10, December 2004.

URL: <<http://library.cern.ch/HEPLW/10/papers/2/>>

Reader Response

If you have any comments on this article, please contact the [Editorial Board](#)

[Top](#)

[Home](#)

[Editorial Board](#)

[Contents](#)

[Issue 10](#)

Maintained by: [HEPLW Team](#)

Last modified: November 2004