Toward conceptual indexing using automatic assignment of descriptors

Arturo Montejo Ráez

CERN - European Laboratory for Particle Physics, Geneva, Switzerland
Data Handling Group

Abstract. Indexing techniques have reached a well maturated state. Digital libraries and other digital collections make an intense use of these algorithms to store and retrieve documents. In the other side, we have browsing techniques, which lets the user to gather the information. Current approaches are not yet advanced enough in order to satisfy the user. At CERN we are working in a indexer based on thesaurus descriptors. With a collection of documents related to thesaurus, user can manipulate them in a more conceptual way. Here we describe the core of this system, the automatic descriptor assigner.

1 Introduction

Indexing techniques has, mainly, focused the attention of *Information Retrieval* (IR from now) researchers, because it was clear that they represents one main problem to be solved and optimized. It is easy to understand such tendency, since indexing has a vast repercussion on the rest of components in an information retrieval system. From older works in IR by Rijsbergen [19] and Salton [16], we can summarize the accessing to the information in this case of use:

1. The user has a specific need of information.
2. The user transmits his/her need of information to the system.
3. The system access the collection and retrieves to the user a set of documents.
4. The user browses the collection and returns a feedback to the system.
5. The system gets feedback from the user so it can perform a better search.
6. The dialog user-system finishes once the user is satisfied with the results obtained.

The well-known *full-text search* is considered as a "philosophal stone" for the implementation of this dialog. We can find several query languages that can enhanced this type of search to enable the user specify more detailed queries. The main problem is the ambiguity of the query, that is, the not trivial task carried out by the user when determining his/her requirements. Some approaches try to solve the problem providing enhanced interface for browsing, and seems that a mixing of good ranking algorithms like *PageRank [13]* together with more intuitive and fast browsing tools, like clustering the result set to make easier the discrimination by the user [14], are in the path to the most suitable solution.

In research environments like CERN, the browsing of documents can be a complex task which involves the gathering on a very large collection. We are working in the developing of more semantic tools which will provide to researchers the ability of jumping from one document to another in a conceptual basis. We use the DESY thesaurus [6] to assign descriptors to High Energy Physics (HEP) related documents . In that way we are adding meta-data which tell us about the semantic content of the document. Since all the descriptors are belonging to a structured thesauri, documents are, therefore, interrelated. We could think in a network in the aim of the *Semantic Web* proposed by Tim Berners Lee et al. [4], but in a very well domain, our digital library.

2 Adding semantic meta-data to the document. Descriptors

Some articles do contain some subject information supplied by the authors (usually only when the journal makes it a condition of publication!). So some journals do have keywords, and quite a few have adopted the PACS classification supported by the American Physical Society [12]. However, these approaches are far from being complete over all documents, so they are not useful for global searching. Therefore, any added data have to be supplied by the creator of the database used for the searching.

This kind of adding of subject material is called subject indexation or keyword[1] enhancement. There are two very different ways of doing this: to choose terms from a fixed thesaurus or to use keywords which can be chosen by the indexer at will. 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 even appear in the text at all, which can give this method a big advantage over any strategy which just uses the text of the document. Examples of fixed thesauri are those used by INIS [1] (International Nuclear Information System, Vienna) and INSPEC [2] (Physics, Computing and Electrical Engineering Abstracts, UK). as well as the DESY Thesaurus.

Many new documents arrive every day at the CERN Library, nearly all of them in electronic form. The task of indexing is mainly performed by indexers working at DESY. Due to the growth in the production of HEP-related papers a new approach to assignment has been developed. Since full-automatic indexers are still far from providing a realistic solution, a computer-based help tool for indexing might be able to be used to in order to ease the work of human indexers.

The *HEPindexer* project intends to propose a preliminary solution, opening the door to research on automatic indexing tools in the area of HEP. This tool proposes descriptors for a given document. In the development of such a system a first step has been achieved: the generation of *main DESY keywords*. These descriptors are generated following a statistical approach [19].

[1] Please note that here, the use of the terms *keyword* and *descriptor* are interchangeable.

```
*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
```

Fig. 1. Extract from DESY thesaurus

Figure 1 shows an extract from the DESY thesaurus. Descriptors labeled with "*" are descriptive (secondary) keywords; those with "-" are non-keywords, while those preceded by a blank are main keywords.

3 Previous work

The availability of large collections of documents in full text format has represented the beginning of a new era in information retrieval. Much research is being done around natural language processing. The early work of Salton [17] provides a good introduction. Many relevant algorithms have arisen for this approach, from classic conflation algorithms to reduce the representation of a document to its essential items (see [15]), to those which treat the document as a whole, identifying discourse trees [11] or conceptual phrases [5].

In the pure sense of descriptors assignment we identify two different tendencies: those ones where the goal is the use of descriptors by humans, and those ones where descriptors are intended to used by other computed tasks. For the first ones we can cite some systems that have been developed during past years, such as BIOSIS, MeSH, the NASA MAI System [10]. For the second use, we have approaches like the probabilistic one of Reginald Ferber [9] and some multilingual approaches like the indexer used in the European Commission [18] for cross-lingual purposes and the MAGIC system of Kutschekmanesch et al. [8].

For us, the use of the descriptors is a mixture of both tendencies. They let us interrelated documents, and they let the user to gather the collection using them. Our system: the *HEPindexer*, is the core of all this, it will propose automatically descriptors for a given full text document.

4 HEPindexer

The algorithm used needs a set of data which must be [3] generated in a *training* process beforehand. Later, this system will be able to propose main descriptors with a reasonable degree of success, as proved through a *testing* process. These two processes require a set of documents as input. HEPindexer is supplied with the training collection of 3.700 documents. This collection was a sample of HEP-related documents and the DESY keywords were supplied for each document. That is, we have a list of documents already labeled by DESY from which our system can learn. After training, we are able to pass a new document to the system and receive as output a list of automatically-proposed descriptors.

4.1 Algorithm

The *training* consists on:

1. Each document is parsed, eliminating *stop words* (articles, prepositions and other words without meaning) and applying a *stemmer* (in order to get the "stem" of each word). Finally, the frequency of every remaining term in the document is computed.
2. For each descriptor, we compute a vector of terms using the following formula:
$$weight(k,t) = \lg \frac{M}{M_t} \sum_d TFIDF_{t,d} \bullet KF_k$$
where
$weight(k,t)$ is the weight of the term t for the descriptor k
M is the total number of descriptors
M_t is the number of descriptors related to term t (that is, the term t appears in M_t documents labeled with k
d is a document
$TFIDF_{t,d}$ is the *document frequency* multiply by the *inverse document frequency* of the term t in document d
$KF_{k,d}$ is the infrequency of the descriptor k for the document d

The *assignment* of descriptors given a new document is performed ranking all descriptors in the thesaurus with a weight computed as follows:

1. The document is parsed, as in the training phase, to get a vector of terms by frequency.
2. The vector is multiply by the matrix of weights between descriptors and terms, given as result a vector of weighted descriptors.

4.2 Results

The system interacts with the user through a web-based interface. Using a web browser, the user can test the system with documents from the test collection or obtain proposed keywords by supplying a new full-text document, either in Postscript or PDF format, or plain text. Although the system can still only propose DESY main descriptors the results are close to 60% in both **precision** and **recall**. That means that an average of almost 60% of the keywords proposed are also contained in the list proposed by DESY, and that almost 60% of keywords proposed by DESY are the same as the returned ones by the system.

This system has now integrated within the CERN Document Server [7]. Improvements 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.

5 Conclusions and future work

Some improvements on the system have to be performed yet. A new HEPindexer is now being programmed based on *Java* and *MySQL*. Other measures will be tested and it is planned to incorporate the capacity of dealing with *multi-words*. After it, the system will be ready to be integrated in a browsing tool for users, providing the feature of gather documents and citations with descriptors from the thesaurus as added values.

References

1. Inis thesaurus. http://www.iaea.or.at/worldatom/publications/inis/inis.html.
2. Inspec thesaurus. http://www.iee.org.uk/publish/inspec/.
3. R. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Acm Press Series, 1999.
4. Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic Web. *Scientific American*, 284(5):34–43, May 2001.
5. Christopher Culy. An extension of phrase structure rules and its application to natural language. Master's thesis, Stanford University, 1983.
6. DESY. The high energy physics index keywords, 1996. http://www-library.desy.de/schlagw2.html.
7. CERN. DH Group, ETT division. The cern document server, 1996. http://cds.cern.ch.
8. Said Kutschemanesch et al. Automated multilingual indexing: A synthesis of rule-based and thesaurus-based methods. In Pub Deutschen Gesellschaft fur Dokumentation, editor, *Information und Markte*, pages 211–224, Donn, Germany, 1998.
9. Reginald Ferber. Automated indexing with thesaurus descriptors: a cooccurrence-based approach to multilingual retrieval. In Carol Peters and Costantino Thanos, editors, *Proceedings of ECDL-97, 1st European Conference on Research and Advanced Technology for Digital Libraries*, pages 233–251, Pisa, IT, 1997. Lecture Notes in Computer Science, number 1324, Springer Verlag, Heidelberg, DE.

10. 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.
11. Daniel Marcu. Discourse trees are good indicators of importance in text. Technical report, Information Science Institute, University of Southern California, 1997.
12. American Institute of Physics. Physics and astronomy classification scheme. http://publish.aps.org/PACS/.
13. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Computer Science Department, Stanford University, 1998.
14. Christopher R. Palmer, J. Pesenti, Raul E. Valdrez, Michael G. Christel, Alexander G. Hauptmann, D. Ng, and Howard D. Wactlar. Demonstration of hierarchical document clustering of digital library retrieval results. In *ACM/IEEE Joint Conference on Digital Libraries*, page 451, 2001.
15. 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.
16. G. Salton. A vector space model for automatic indexing, 1975.
17. Gerard Salton. Automatic text analysis. Technical Report TR69-36, Cornell University, Computer Science Department, June 1969.
18. Ralf Steinberger. Cross-lingual keyword assignment. In L. Alfonso Ure na López, editor, *Proceedings of the XVII Conference of the Spanish Society for Natural Language Processing (SEPLN'2001)*, pages 273–280, Jan (Spain), September 2001.
19. C. J. van Rijsbergen. *Information Retrieval*. London: Butterworths, 1975. http://www.dcs.gla.ac.uk/Keith/Preface.html.