Extracting Keywords from Digital Document Collections

Anna Jonsson Swedish Institute of Computer Science Human Computer Interaction and Language Engineering Laboratory Kista, Sweden E-mail: anjo@sics.se

Abstract

An indexing tool was built to provide for one of several information seeking tasks. In accordance with the basic principles of work held by the HUMLE laboratory at SICS, a solution regarding indexing would be a semi-automatic tool. This approach is also relevant as the continuation of the indexing project is conducted in co-operation with the Swedish Parliament, where a staff of professional indexers currently is investigating the utility of automatic and semi-automatic indexing tools to raise productivity.

1 Introduction

Digital libraries are complex information systems, which augment and extend traditional libraries by affording users better support for human problem solving and problem formulation. Digital libraries should be understood to be more than a haphazard collection of electronic resources and associated technical widgets for creating, searching, and using information in various media and over networks. They are, or should be, tailored to the needs and tasks of a group or several groups of users, and their functional capabilities should support the information needs and uses of those individuals and groups.

Digital libraries are both an extension and integration of existing information sources, and through the advent of new technology and adjustment of tried and familiar technologies, a completely new concept. While digital libraries typically improve certain aspects of traditional libraries, most often today they leave other aspects unaddressed, which will decrease their usefulness. Traditional information institutions not only make information resources available to the public, but actively select, collect, organise, and preserve them, engaging in numerous behind-the-scenes tasks seldom addressed, or taken for granted in their digital counterparts.

Despite recent advances in both computer technology and computational linguistics, retrieving and extracting useful information in large document collections is still very troublesome. Freetext search is certainly useful and fast, and generates a generous amount of results, but distinguishing the relevant documents from the non-relevant in the abundance of returned documents is a problem. Other systems for structuring information to enhance availability has traditionally been by storing information about

documents, books, and texts in bibliographic cards; and by indexing the documents by lists of keywords or keyphrases.

2 SICS Digital Library

To understand the context to our current indexing work, this chapter gives an overview of the theoretical background important in the Digital Library Project at SICS (DigLib), as well as a short description of the project itself.

2.1 Theoretical Background

The belief that one kind of information retrieval system, i.e. freetext search systems, can suffice for, and even replace other systems, and thus provide for all information seeking needs users have, is widely common today. This maybe due to the recent technological advances that have solved a considerable amount of the problems this area suffered initially, mainly regarding speed and storing possibilities. However, Belkin and others have analysed user strategies for information seeking (Belkin & Cool, 1993; Belkin, Cool, Stein & Thiel, 1995), and recently Belkin and Carballo (1998), based on extensive user observations, found that humans utilise a multitude of strategies in the task of information seeking. The four strategies people spend most time on are: finding a known information object; recognising useful information objects by scanning through an information resource; evaluating the usefulness of information objects; and determining the content or structure of a collection of information objects. The first of these strategies, "finding a known information object" can be seen as corresponding to bibliographic cards, containing information about author, title, publishing data, and possibly an abstract, used for example in traditional libraries. An example of what "recognising useful information objects by scanning through an information resource" means, can be the behaviour we adopt when browsing through web pages. The third strategy we spend time on, "evaluating the usefulness of information objects", is e.g. when we try to distinguish what documents are relevant to our query after a search, for example a search on Altavista. And finally, "determining the content or structure of a collection of information objects" is what we do when familiarising ourselves with a book by looking at the table of contents, or by looking up the keywords listed in the book's index.

So, the conclusion is that multiple information seeking strategies need to be met by several information seeking tools, which clearly indicates that a single information access tool would not be sufficient.

2.2 DigLib

SICS runs DigLib, a project for the study and application of digital libraries. The central issue for DigLib is focusing on usage: studying how and why users interact with document collections, and trying to build tools incorporating new technology to aid users in the tasks we find they try to solve. The leading principle is that general solutions to information access problems tend to be unsatisfactory, and that tailoring technology to

specific requirements of professional users is more rewarding than trying to automate general tasks for all.

To provide for the above-mentioned information seeking strategies proposed by Belkin and Carballo, a platform including different tools was built within DigLib. As a tool for "finding a known information object", we included Dienst, a bibliographic look-up engine. Dienst performs a search in bibliographic records, where, among other things, title and author are specified. It provides rapid, consistent, and predictable results: it indexes documents by a small number of highly relevant fields without bothering with the full document texts (Davis & Lagoze, 1994). "Recognising useful information objects by scanning through an information resource" was made possible through a systematic structure of HTML links, organised hierarchically by information source. The third prototypical information seeking interaction "evaluating the usefulness of information objects" cannot be met by adding a single tool, as this concerns different aspects of individual documents or sets of documents, and requires specific techniques for information refinement. Evaluating a document in regard to a user's information need, can for example be accomplished by comparing the document to other documents in the collection or by custom-made summarisations of retrieved documents.

There are several ways to support the fourth information seeking interaction style, "determine the content of a collection of information objects". Although graphical visualisation techniques using various metaphors show promise of usefulness, other solutions must be offered, since up to 64 percent of the population have difficulties using maps for orientation even in straightforward navigation tasks (Streeter & Vitello, 1986). We implemented the Keyword Extraction Function (KEF, see further in section 3.1 below), that extracts and presents keywords from documents in a way that resembles an index in a book - something most of us are familiar with, and many of us use for precisely the purpose of familiarising ourselves with the contents of a book.

3 Previous Work

Within the DigLib project we have, as outlined above worked on integrating several tools that meet different information seeking strategies. The tool we developed in-house corresponds to one of these strategies, and is described below.

3.1 The Keyword Extraction Function

The tool called the *Keyword Extraction Function* (KEF) was the first prototype to an indexing tool. The function takes all words from the text-files and applies a lexical filter, which selects all nouns from these texts. For this it is dependent on a part-of-speech tagger. We investigated the possibility of using other criteria for keyword spotting, such as word length; long words tend to be topic-specific, especially in a compounding language such as Swedish. The results were not completely discouraging, but we found that the benefits a tagger gave were not limited to term spotting, but included conflation of morphological variants - which in general is desirable. We concluded that tagging is necessary for term extraction.

Nouns were chosen as index words, as they seem more information dense than other word-classes, and are most often used in retrieval queries (Källgren, 1984; Källgren, 1992). Complex noun phrases were disregarded, after having analysed a sample set of documents. In addition Ingwersen (1992) states that automatic indexing techniques based on single words are quite effective, and multi-stream information retrieval experiments shows that single term retrieval in general is the single most effective knowledge source for information retrieval (Strzalkowski et al., 1997), compared to, among others, multi-word terms.

By selecting all nouns occurring in the texts for further inspection, KEF overgenerates terms. The assumption was that overgenerating terms and excluding non-relevant ones is a safer bet than attempting to pick only the most relevant ones. A number of researchers have in various ways shown that frequency or repetition in addition to lexical category is important for modelling term relevance (Luhn, 1959; Salton, 1989; Justeson & Katz, 1995).

Our conclusions from this work were that morphological and lexical tagging is necessary for term extraction and that a keyword index needs further techniques for refining the choice of included terms. Current experiments include statistically based term selection metrics, and the possibility of generating different index term lists for different purposes: a shorter list could conceivably be used for overview; a more exhaustive list for finding precisely which items to peruse further.

4 Current Work

There are today many organisations that daily deal with large amounts of documents, index them manually, i.e. create lists of keywords describing the document, in order to enhance the documents availability for information finders within the organisation as well as, in some cases, for the general public. Indexers sometimes have access to knowledge databases of some sort, often a structured ontology or knowledge model to aid them in their work. The quality of these manually produced indexes is high, how-ever, although when having access to ontologies, humans do not seem to index in a consistent way (Earl, 1970, Kowalski, 1997), and the work is both time-consuming and expensive. As comparison, automatic indexing, which has undergone a dramatic change to the better since the beginning in the 50s and 60s, gives quick results that can be said to be consistently done, and it is far less expensive than manual indexing, but the quality is still questionable (Salton, 1989, Kowalski, 1997).

So, bearing in mind that manual indexing is of high quality but time consuming and inconsistent, and automatic indexing is fast, consistent but of lower quality - how can we improve the quality of indexing, and relieve pressure on those who are professional indexers, and increase productivity?

4.1 Principles of Work

As mentioned above, the leading principle when working with DigLib, was to avoid general solutions. In our case, and as an answer to the question ending the previous section, the solution to us is semi-automatic tools. To understand why, a description of some of the basic principles important for the work done at the HUMLE laboratory might be necessary. These principles concern: system context; methods for design; and the development of tools for professionals.

Firstly, it is important to adapt a system to an existing situation, understanding the language used in some specific context, and to include modality-specific information in the system specification. There is often no best general solution available, but a system needs to be customised to fit in the workflow of the organisation. Secondly, it is important to bear in mind the difficulty involved when trying to collect and collate the right sort of information; information that has bearing on the design process. That is, finding a method of relating information about a workplace and the individuals therein to the design process is essential. And thirdly, the professional user shall be aided in the task at hand, and it should be an intelligent aid system. Complete automation should be regarded with some degree of scepticism. The goals are to raise productivity, efficiency, and quality of information work.

4.2 The Swedish Parliament

One of the organisations we co-operate with regarding indexing is the Swedish Parliament. They index large numbers of documents yearly, in order to make them accessible both to information specialists and to the general public, and their work on this plays an important societal role. This work has progressed for a long period of time, during which they have developed an extensive hierarchically organised domain specific thesaurus (or knowledge base). Figure 1, below, shows an excerpt of the thesaurus: the word *arbetshandikapp* (work disablement) and it's broader term (BT), narrower terms NT, related terms (RT) and a description of what the word means (SN).

Arbetshandikapp BT Arbetsliv

```
NT Arbetsiiv
NT Arbetsbiträde
NT Näringshjälp
NT Skyddat arbete
RT Anställningsfrämjande åtgärder
RT Handikapp
RT Lönebidrag
SN Nedsatt arbetsförmåga pga fysiska, psykiska, förstånds-
mässiga eller socialmedicinska handikapp – däri inbegripet
missbruk av alkohol eller annat berusningsmedel.
```

Figure 1. Excerpt from the thesaurus developed at the Swedish Parliament.

As a consequence of the parliament's long history of manual indexing, radical departures from tradition is very much undesirable. Meaning that an aiding indexing tool in essence must function as a re-implementation of the tasks currently performed.

4.3 Towards a Domain Specific Indexing Tool

The development of the new indexing tool takes a slightly different starting point than KEF. The input comes in the form of a list of keywords generated by means of standard tf*idf calculations. As before, we have concentrated on nouns as keywords, at least as an outset. The new approach, from our point of view, is making use of the thesaurus, (mentioned above), for shortening the list of keywords. By aggregating several occurring narrower terms, the tool can then suggest to the human indexer, with some measure of confidence, a broader term that describes the document on a higher level of abstraction. For example, if the list of keywords shows that the document contains several occurrences of specific types of banks (e.g. Affärsbanker, Föreningsbanker, Investeringsbanker, Sparbanker) the indexing tool will suggest their broader term *Finansinstitut* (Finance Houses) to describe the document. The confidence measure will reflect the coverage of the terms occurring in the document that are corresponding to the terms listed in the thesaurus.

In order to improve the performance of the indexing tool, automatic evaluation based on previous manual indexing will be implemented. The manual indexing done at the Swedish Parliament over many years has been stored and used for an enhancement of the indexing accuracy. This will be taken into account also for our indexing tool.

Another facet that will be taken under consideration in the further development of the tool is the temporal aspect. Term meanings change over time, and index terms shift over time. This should be made visible to both the indexer and the viewer/user of the finished index. When e.g. looking up a new term, it may be relevant to also retrieve documents covering the same concept, although it might be described using slightly different terms.

In the future we want to enable the indexing tool to recognise when new terms should be added to the thesaurus. If there seems to be a new term frequently occurring in a number of documents indexed with the aid of the tool, it should make the manual indexer aware of this, by recommending that the term be added to the thesaurus.

5 Discussion

This article has given the settings to the current work performed within the DigLib project at SICS. The theoretical background is that people use various strategies when seeking information (Belkin & Carballo, 1998), and the strategy we have focused on is one way of allowing users to familiarise themselves with an information resource. This is accomplished by presenting an index of the document. The work started with the Keyword Extraction Function, implemented in SICS' Digital Library platform (Hulth & Jonsson, 1999), and has later on developed into the semi-automatic indexing tool currently under development.

Semi-automatic indexing may not be too hard to accomplish per se, as it is a rather straightforward task. Using the semi-automatic indexing in combination with ontologies may, however, be more of a challenge, especially the question of how to combine two or more differently structured knowledge models, and allow them to communicate and co-operate with each other. This will be one focus of research that we will be experimenting with in the near future.

Another focus is the question of what makes a good index. There is more to indexing than merely making lists of words, whether they have frequency measures and weights of various kinds attached to them or not. What criteria do humans use when indexing a document for example?

When we have realised the implementation of the tool, and evaluated its performance in the authentic environment, adapted to the requirements of the workplace, we will know to what extent our presumptions regarding semi-automatic indexing were correct.

Acknowledgements

A lot of the previous work was performed in co-operation with Anette Hulth.

References

- Belkin, N.J. & Carballo, J.P. 1998. Understanding and Supporting Multiple Information Seeking Strategies, a TIPSTER Phase III Research Project. URL: http://www.scils.rutgers.edu/tipster3/18month/index.htm
- Belkin, N.J., Cool, C., Stein, A. & Thiel, U. 1995. Cases, Scripts and Information Seeking Strategies: On the Design of Interactive Information Retrieval Systems. Expert Systems with Applications.
- Belkin, N.J., Marchetti, P.G. & Cool, C. 1993. BRAQUE: Design of an interface to support user interaction in Information Retrieval. Information Processing & Management Vol 29, 3, 325-244.
- Davis, J.R. & Lagoze, C. 1994. A Protocol and Server for a Distributed Digital Technical Report Library. URL: http://cs-

tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrl.cornell/TR94-1418?a

- Earl, L.L. 1970. Experiments in Automatic Extracting and Indexing. Information Storage & Retrieval. Vol. 6, pp. 313-334. Pergamon Press.
- Hulth, A. & Jonsson, A. 1999. An Experimental Digital Library Platform A Demonstrator Prototype for the DigLib Project at SICS. Stockholm. SICS Technical Report T99:02.
- Ingwersen, P. 1992. Information Retrieval Interaction. Taylor Graham Publishing.
- Justeson, J. & Katz, S. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering 1 (1) 9-27*. Cambridge University Press.
- Kowalski, G. 1997. Information Retreival Systems: Theory and Implementation. Kluwer Academic Publishers.
- Källgren, G. 1984. Automatisk exerpering av substantiv ur löpande text. Ett möjligt hjälpmedel vid datoriserad indexering? IRI-rapport 1984:1. Institutet för Rättsinformatik. Stockholm University. (In Swedish.)

- Källgren, G. 1992. Making maximal use of surface criteria in large-scale parsing: the MorP parser. Papers from the Institute of Linguistics. (Publication 60). University of Stockholm.
- Luhn, H.P. 1959. Auto-Encoding of Documents for Information Retrieval Systems. In: Boaz, M. Modern Trends in Documentation. (Ed.). London: Pergamon Press. (Reprinted in: Schulz, C.K. (1968). H.P. Luhn: Pioneer of Information Science, selected works. (Ed.). New York: Sparta.).
- Salton, G. 1989. Automatic Text Processing. Addison-Wesley Publishing Company.
- Streeter, L. & Vitello, D. (1986). A Profile of Drivers' Map-Reading Abilities. Human Factors, 1986, 28(2), 223-239.
- Strzalkowski, T., Guthrie, L., Karlgren, J., Leistensnider, J., Lin, F., Perez-Carballo, J., Straszheim, T., Wang, J. & Wilding, J. 1997. Natural Language Information Retrieval: TREC-5 Report. *Proceedings of the fifth Text Retrieval Conference*, Donna Harman (ed.), NIST Special Publication, Gaithersburg: NIST.