The Feasibility of Using the Semantic Components Model for Indexing Documents in Digital Libraries

Susan Price[1] **Marianne Lykke Nielsen**[2] **Lois Delcambre**[1]

[1]Department of Computer Science, Portland State University, Portland, Oregon
[2]Royal School of Library and Information Science, Aalborg, Denmark

Introduction

Finding one or more documents that exactly answer a targeted information need often fails, especially in digital libraries that lack the hyperlink structure that is so successfully exploited by the page rank algorithm [1]. In the absence of extensive hyperlinks, successfully matching document requests to document content is essential. Assignment of keywords using human intellectual processes is expensive and prone to inconsistency [2, 3]. Automated full-text indexing is less expensive but requires the searcher to anticipate the language used in relevant documents [4]. We have developed a new model, which we call *Semantic Components*, that leverages expert knowledge about how information is organized and expressed within the domain and is intended to facilitate precise searching in domain-specific libraries [5]. We hypothesize that semantic component indexing will yield improved search results over automated full-text indexing and that indexing using this model will be faster (and therefore cheaper) and more consistent than human keyword assignment.

Research Overview

In order to fully evaluate the feasibility and usefulness of the Semantic Components model, we have four main areas of inquiry: (1) How easily can the model be applied to a particular document collection? (2) How easily can searchers use the model to represent information needs? (3) How easily can documents be indexed using the model? (4) Is the model useful for retrieving documents? This presentation will focus on the third question, regarding indexing with semantic components. In this paper we briefly introduce the model and its uses, then describe our methodology for an indexing experiment that will be completed in June 2006. Our presentation at the NKOS workshop will focus on the results of that experiment.

Semantic Components Model

The Semantic Components model consists of document classes and semantic components. Document classes are classifications of documents based on factors such as type of topic and intended audience. For example, a digital library containing documents for physicians might contain three document classes: *document about a disease*, *document about a diagnostic or therapeutic procedure*, and *document about a drug.* In the collections we have analyzed, the classes correspond to the notion of document genres in that they have a specific purpose and are part of an organizational communication process. Semantic components (SCs) are important aspects of the main topic that commonly appear in documents of a particular class. Each class is associated with a small set of SCs. For example, documents about diseases might contain information about *causation*, *prevention*, and *treatment* (three SCs) while documents about procedures might contain different SCs, such as *indications* and *risks*. A semantic component instance is one or more segments of text, not necessarily contiguous, that contains information about a particular aspect of the topic of a document. For example, the text that describes

treatment options in a document about colon cancer (an instance of the *document about a disease* class) is an instance of the *treatment* SC. An SC instance may be associated with a structural element in the text that helps with identification of the instance, but the model does not require such a structural indication. Figure 1 shows SC instances in a document.

There are two main ways that information retrieval systems can exploit SC information. First, a search result that displays the SCs provides a synopsis of document content that can help a searcher assess likely relevance and choose which documents to view. Second, a query language can incorporate SC information, allowing a searcher to specify which SCs are desired, or to apply the query string (text words, phrases, or keywords) to a particular SC. The SC specification can be used as either a filter or a ranking parameter. Directing the full-text search to specific SC(s) may facilitate precision while minimizing the effect on recall. Figure 2 illustrates a search specification.

Critical to the usefulness of the SC model is being able to identify SC instances with a reasonable degree of consistency using a reasonable amount of resources. An important feature of the model is that the sets of document classes and SCs are identified in advance for a particular document collection. Indexers will need only to choose the document class then identify the presence and location of SC instances. Instead of determining what concepts to index and what term(s) should represent each concept, an indexer will decide whether or not a particular segment of text pertains to each of a small set of aspects (the SCs) of the document topic. This may reduce the cognitive demand of indexing and facilitate speed and consistency.

Indexing Experiment

The indexing experiment we will perform in June will compare keyword indexing to SC indexing with respect to speed, consistency, quality, and perceived difficulty. Specifically, the study will address the following questions: (1) Can human indexers identify SC instances more quickly than they can assign indexing keywords? (2) Is identification of SCs more consistent than assignment of keywords? (3) How do keywords assigned by indexers and SCs identified by indexers compare to a reference standard generated by a consensus of indexing and domain experts? (4) Do indexers find the identification of SCs to be easier (less cognitively demanding) than assignment of keywords?

Twenty subjects who currently index documents for sundhed.dk, the national Danish health portal, will each index nine sundhed.dk documents using both keywords and SCs. The keyword indexing task will mimic the indexers' current indexing process, in which indexers may assign terms from two domain-specific vocabularies plus "free" terms (terms not in any of the controlled vocabularies). The SC indexing task will consist of choosing the document class and marking the location of instances of SCs. We will supply the indexers with a list of document classes and their corresponding SCs. Both indexing tasks will be performed using pens and paper documents in order to isolate the intellectual indexing tasks from the confounding influence of computer application interfaces that would necessarily be different for the two tasks. Subjects will be asked to identify the location and extent of SC instances by circling and labeling segments of text with colored pens. We refer to the SC identifications as SC markup.

Measurements will include the time required to index documents using each technique, the consistency and quality (in relation to subject analysis and translation to keywords) of keyword indexing, the consistency of document classifications and SC markup among the group of indexers, and the similarity of the classifications and SC markup to the reference standard. One of the challenges of this work is determining the best technique for comparing SC markup. SC indexing does not exactly fit any of the previous models of indexing or related IR tasks, so existing evaluation methodologies cannot be applied directly. We treat an SC markup as a set of binary text classification tasks in which the SCs are classes and each word in the document is classified as belonging, or not belonging, to an SC instance. This allows us to generate measurements that reflect the similarity of SC markup.

In addition to the quantitative measurements we will use questionnaires and interviews to assess the indexers' attitudes regarding the two indexing techniques, especially with respect to perceived cognitive ease and confidence in their indexing choices. We will also determine whether these results are influenced by the subjects' amount of experience or education related either to indexing or to the medical domain.

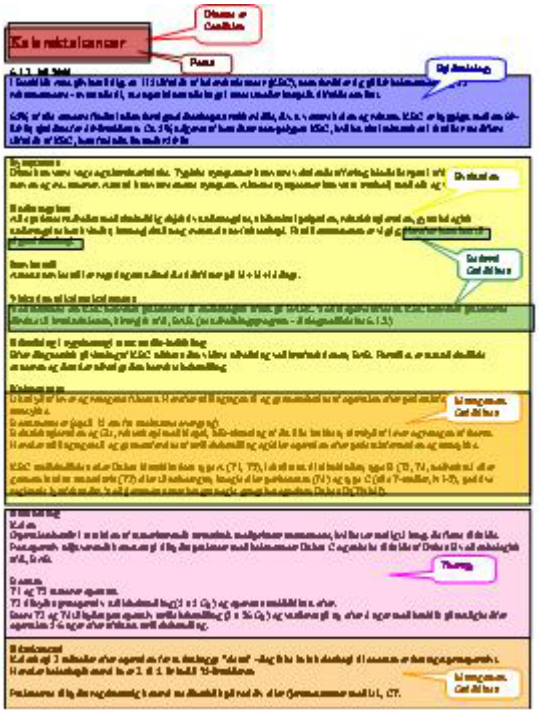

Figure 1: Danish health document (excerpted from http://www.sundhed.dk) with various semantic components marked.

Specifying a search

Primary focus of search:
kolorektalcancer
Search for:
sigmoideoskopi
In:

- [x] Evaluation
- [ ] Therapy
- [ ] Management Guidelines
- [ ] Referral Guidelines
- [ ] Prevention
- [ ] Risk Factors
- [ ] Prognosis
- [ ] Etiology
- [ ] Associated Conditions
- [ ] Epidemiology

Figure 2: A search specification

References

1. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, pp. 107-117, Brisbane, Australia, 1998.
2. F. W. Lancaster. *Indexing and Abstracting in Theory and Practice*. Third ed. University of Illinois Graduate School of Library and Information Science: Champaign, IL, 2003.
3. Mark E. Funk and Carolyn Anne Reid. Indexing Consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71(2), pp. 176-183, 1983.
4. J Rowley. The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, 20(2), pp. 108-119, 1994.
5. Susan L. Price, et al. Using Semantic Components to Facilitate Access to Domain-Specific Documents in Government Settings. In *The 7th Annual International Conference on Digital Government Research (dg.o)*, San Diego, California, May 21-24, 2006.