How doctors apply semantic components to specify search in work-related information retrieval

Marianne Lykke
Aalborg University, Denmark

Susan L. Price
Microsoft Corporation, USA

Lois L. M. Delcambre
Portland State University, USA

Abstract

Workplace searching is often context-specific and targets a ’right answer’ within some domain-specific aspect of the search topic. We have developed the semantic component (SC) model that allows searchers to specify a search within context-specific aspects of the main topic of documents. The goal of our study was to gain insight into how family practice physicians at sundhed.dk, a national healthcare portal in Denmark, applied the SC model to formulate queries to solve work-related search tasks. The results showed that doctors used the model purposively when choosing search facets and search concepts. They were relatively consistent in their use. The findings provide promising evidence of the model’s potential usefulness.

Introduction

Studies of workplace search demonstrate that workplace searching is different from general searching Professional workplace queries are typically targeting a single ‘right answer’, and tend to result in a small set of useful documents (Fagin et al., 2003). The correct answer is context-specific and often limited to domain-specific aspects such as information characteristics, communication channels, document status, and subject categories, e.g. products, technology, customers (Freund, Toms & Waterhouse, 2005). We have developed the semantic component (SC) model, which supports indexing and search, that allows searchers to structure and specify the search within context-specific aspects of the main topic of the documents. We have evaluated the SC model in a realistic, interactive searching study with family doctors using the web-based national health portal Sundhed.dk intended to support work-related information search tasks during a primary care visit (Price, Nielsen, Delcambre, Vedsted & Steinhauer, 2009). The study showed an improvement in document ranking of 35.5% as measured by MAP (Mean Average Precision) and 25.6% by nDCG (normalized Discounted Cumulative Gain) when SC indexing and searching was added to a standard search engine.

In the present paper we report on how doctors applied the SC model when formulating search queries. Our goal is to gain insight how they used SCs to express the facets of the search task. Altogether, we want to extend our understanding about the SC model in order to guide the design of information retrieval systems.

We specifically address the following research questions:

- How did the doctors use SCs to formulate queries to solve search tasks?
- How did they express the search facets: by SCs, search terms or search filters?
- Which SCs did they use?
- Did they use the SCs consistently?

The remainder of the paper is organized as follows. Section 2 presents the SC model and the associated search interface. Section 3 describes the research design. Section 4 presents and discusses the results. Section 5 concludes and presents propositions for further research.

The semantic component model

The SC model is developed in order to facilitate formulation of structured queries that can use domain-specific aspects of documents. The model divides a given collection into a set of document classes, where each class has an associated set of SCs. Documents may be classified by a variation of characteristics, e.g. by type of topic, by purpose of the document, by form, etc. The appropriate classification depends on the nature of the document collection. In a health-related collection topic type is a natural axis for classification. Doctors search for information about diseases, drug dosing, causes of symptom X, management of disease or finding X (Ely et al., 1999). In the case of sundhed.dk we have found documents that correspond to these topics, e.g. documents about diseases (one document class), documents about medications (a second document class), and documents about treatment methods, e.g. chemotherapy (a third document class). For each class we have identified a finite set of aspects of the topic that are important in the domain and reflect a general understanding of the type of information needs expected. For example, the document class about diseases often contains information about *symptoms* and *treatment* whereas documents about medications often contain information about *dosage* and *side effects*. These aspects constitute the *semantic components* of the document class; we define an SC instance as one or more text segments about a particular aspect of the main topic of the document. The segments of text that comprise an SC instance can vary in length and may or may not be contiguous. Each SC instance consists of segments of text that may not correspond to structural elements in the document. Table 1 shows the document classes and semantic components used in the study to index and search. The SC model allows the user to specify the search in two ways: 1) a searcher can specify that his search is for documents about a core topic (described in terms of the document class) that also contain a particular SC indicated by an *, and 2) the searcher can search for terms within SC text segments.

Document Class	Semantic Components	Document Class	Semantic Components
Clinical problem (e.g. diseases)	General information Diagnosis Referral Treatment	Clinical unit (e.g. hospital)	Function and specialty Practical information Referral Staff and organization
Clinical method (e.g. chemotherapy)	General information Practical information Referral Aftercare Risks Expected results	Drugs (e.g. specific medication)	General information Practical information Target group Effect Side effects, interactions and contraindications)
Services (e.g. free psychological help covered by public insurance)	General information Practical information Referral	Notice (e.g. call for patients for clinical trials)	General information Practical information Qualification

Table 1. sundhed.dk SC model

Document classes are related to the familiar notion of document genre. A number of authors have suggested using document genre to improve information retrieval. Crowston and Kwasnik (2003) argued that communication of document types and structure can be useful specifically for users with domain-specific information needs as an intermediating mechanism to connect information needs and information. Dillon (1991) and Bishop (1999) provided empirical evidence that readers can manipulate sub-document components in the context of a familiar document model, but did not specifically address information retrieval uses. Freund, Toms and Waterhouse (2005) analyzed the information tasks of software engineers and demonstrated a correlation between information tasks and document genre. Turner et al. (2005) created a model of documents in the public health domain in which genre was one component. They used content analysis and expert users to identify elements in public health gray literature that could be extracted to create a searchable database of document surrogates. Some of the key elements in the proposed surrogate, such as description of the problem, description of the intervention, and target population are similar to semantic components but were not linked to particular document types. The SC model builds on this assumption that domain experts know the genres or document classes within a certain domain.

SC instances are similar, but not necessarily identical, to the classical facets used within library and information science to structure a subject area. Vickery (1960) defines facets and facet analysis as the sorting of terms in a given field of knowledge into homogeneous, mutually exclusive facets, each derived from the parent universe by a single characteristic of division. A subject field has innumerable characteristics, and may be divided into a variety of facets. Aitchison, Gilchrist and Bowden (2002) list the following 'fundamental' facet categories that may be applicable in many subject fields: Entities, Attributes, Actions, Space and Time. Some SCs correspond quite naturally to facets. For example, segments that contain information about *treatment* in a document about diseases may be considered as an action facet. However, SCs can also contain information that may not be considered a facet, such as *general information* that combine two or more concepts that are different aspects of a topic, such as combining epidemiology and natural history about a disease. SCs are used to specify which kind of information is given about a document's general topic, and may specify sub-topics, type of information or target group, depending on the domain-specific collection. Since SCs are intended to facilitate retrieval, and not intended to describe the domain, knowing the contents of a particular document collection, or the common information needs among users of the collection, may lead to selecting SCs with varying degrees of specificity to represent document content.

Methodology

The interactive searching study was an experimental comparison study between a baseline system and an experimental system. At the time of the study the test systems contained nearly 25,000 documents. The baseline system used a combination of human, controlled indexing and automatic, full-text indexing, and the experimental system had the same features plus the ability to further specify the search using the SC model. The controlled indexing terms in both systems were from three health-related controlled vocabularies.

The experimental system includes the SC model and provides a simple search box plus two search filters that are controlled by pull-down menus, one to filter documents by the region to which the documents apply, and one to filter documents by an existing sundhed.dk document classification labelled Information Type. To search the system, the searcher types one or more terms into the search box and optionally chooses an item from the pull down menus for the two filters. In addition, the searcher can enter one or more terms into one or more of the text boxes for the semantic components to search the terms within SC text segments, or enter an * to restrict the search to documents that contain a particular SC.

Each experimental session consisted of four search sessions. Each searcher used the baseline system for two consecutive tasks and the experimental system for two consecutive tasks. Each task represented a typical information search activity that might be encountered in the context of a patient visit. Table 2 provides a condensed summary for each search task with an indication of search facets. The order of tasks and system use were randomized in order to avoid any effect that the order might have on the result, because the test subjects might gain new IR and subject knowledge during the search. Another intention was to neutralize any effect caused by increasing familiarity with the experiment.

Search task	Topic description	Reference standard
A	Ex-smoker, cough, fatigue, shortness of breath. Two tentative diagnoses: 1) Chronic obstructive pulmonary disease (COPD) and 2) Lung cancer. First step is examination of COPD due to waiting time for lung cancer examination. Find documents that help you to decide the referral procedure to follow to evaluate for COPD	Facets: Search concepts Diagnosis: Chronic obstructive pulmonary disease (COPD) Admin activity: Referral Health care activity: Evaluation Information type: Referral guidelines Region: Aarhus
B	Pregnant woman (week 23), light bleeding, no pain, no other abnormal conditions. Find documents that help you to decide if and how the patient should be referred to acute examination at hospital	Facets: Search concepts Condition: Pregnancy Diagnosis: Spontaneous abort Phase: 2 Trimester Symptom: Bleeding Admin activity: Referral Information type: Referral guidelines; Treatment description Region: Aarhus
C	Woman with two un-intentional miscarriages, ready to get pregnant again. Find documents that help you to decide if the patient should take folate, and at what dose	Facets: Search concepts Condition: Pregnancy after two un-intentional miscarriages Health care activity: Drug prescription Mechanism: Folate Information type: Treatment description
D	Man attacked with knife. Now nervous and afraid to leave his apartment. Find documents that help you to decide if the patient can be referred for free psychological help (covered by public insurance)	Facets: Search concepts Diagnosis: Anxiety Treatment: Psychological treatment Health service: Public insurance Admin activity : Referral Information type: Referral guidelines

Table 2. Information search tasks

We collected data about querying behaviour through search logs, questionnaires, verbal protocols, interviews, and relevance assessments. In the present paper, we focus on search log data to investigate how SCs were applied. A query is defined as the enquiry submitted on one occasion to the search engine as a combination of 1) one or more search terms put into the search box, 2) a choice of search filter items from the pull-down menus, and 3) one or more search terms or * entered into the text boxes for the SCs. In order to get an understanding of the queries, we conducted a facet analysis of the search terms put in the search box and categorized the search concepts used to express the search facets. We compared the result to an ideal reference standard in order to get an understanding of how the searchers chose to express the search facets (for the reference standard see Table 2). The facet analysis and development of the ideal reference standard were carried out in collaboration with two experts in facet analysis and information retrieval. First the experts made the analysis separately, next the results were compared, and final differences were solved by group discussion. Two members of the research team coded the queries. Differences were discussed and jointly solved.

Results and discussion

The variables and results of the study are presented in Table 3.

Variable	Definition and measurement	Result
Queries per session (average)	Average number of queries per session	3.2
Sessions with reformulations	Percentage of sessions with reformulations	65.0%
Search concepts per query	Average number of search concepts per query	1.5
Search concepts per session	Average number of search concepts per session	4.7
Queries with search terms	Percentage of queries with query terms	100%
Queries with region filter	Percentage of queries with 'region' filter	34.0%
Queries with information type filter	Percentage of queries with 'information type' filter	29.3%
Queries with SC 'term', first specification (SC1term)	Percentage of queries with SC1term	19.9%
Queries with SC '*', first specification (SC1*)	Percentage of queries with SC1*	48.2%
Queries with SC 'term', second specification (SC2 term)	Percentage of queries with SC2 term	3.7%
Queries with SC '*', second specification (SC2*)	Percentage of queries with SC2*	4.2%

Table 3. Variables of the study

The doctors used 281 search concepts to express the search facets, on average 1.5 concepts per query and 4.7 per session. They used the region filter in 65 queries (34%) and the information type filter in 56 queries (29%). They used SC1* in 92 queries (48%) and SC term in 38 queries (19%). As the second specification SC2, they used SC2* in 8 queries (4%) and SC2 term in 7 queries (4%).

The searchers did not search for all search facets from the reference standard, but they were relatively consistent in their choice of search facets and concepts. For task A, they focused on the diagnosis facet. They expressed the facet by the search term *COPD* (chronic obstructive pulmonary disease) in all queries except one. In 58% of the queries they further specified the search to the diagnosis SC in the clinical problem document class. In 50% of queries they limited their search using the region filter and in 42% using the information type filter. For task B they searched for a combination of *pregnancy* (condition facet) and *bleeding* (symptom facet) in 58% of the queries. In 49% of the queries they filtered by region and by information type in 18% of the queries. In 65% of the queries they specified the clinical problem document class and specified further to the referral SC (23%), treatment SC (16%) or diagnosis SC (11%), either for terms or *. For task C they searched for a combination of *Folate* (mechanism facet) and *pregnancy* (condition facet) in 63% of the queries, and filtered by region in 19% and information type in 23% of their queries. They most consistently specified the clinical problem document class, together with the treatment SC* in 19% of the queries and with the SC term in 13% of the queries. In other queries, a variety of document classes and SCs were used. For task D they searched for *psychology* (treatment facet) in combination with either *referral* (admin activity facet) or *treatment* (treatment facet). They filtered by information type in 44% of the queries and region in 22%. In 38% of the queries they limited the search to the referral SC, mostly in the service document class, but also in the clinical problem, clinical unit, and clinical method classes.

Most searchers started the searching task using these core facets and concepts. Only in a few cases did the searchers chose to express the search topics by other facets, e.g. for task B they used the diagnosis facet and the search term *spontaneous abortion* instead of the symptom facet *bleeding*, hereby interpreting and translating the search problem from facts (bleeding) to diagnosis (spontaneous abortion).

Other facets were used for modification and reformulation. For task A they reformulated by adding the information type filter. For task B they reformulated by using the diagnosis facet (*spontaneous abortion*), the phase facet (*2. Trimester*), or used another symptoms facet (*heart sound and movement*). For task D, they reformulated by adding the diagnosis facet (*anxiety*). They primarily reformulated by adding the information type filter for task C. In fewer cases they reformulated by using synonym variants. Neither hierarchical broader terms nor narrower terms were used. The searchers chose search terms at the same level of specificity as the task description.

In general, the doctors used search terms to formulate the core clinical problem, e.g. *COPD, bleeding pregnancy, folate pregnancy,* and *psychological treatment*, whereas they used the SCs to express the domain-specific angle of the problem, e.g., how to diagnose and refer a COPD patient, how to refer a bleeding pregnant woman, how to prescribe folate, and how to refer a patient for free psychological help. They mostly used SC* thereby limiting the search to documents that contain the selected SC instance instead of using an SC term that limits the search to look for specific terms within SC instances. However, they also used SCs to combine and specify the clinical concepts, e.g. they searched for the term *folate* in the treatment SC to specify that they were searching for vitamin folate as a kind of medication. They were fairly consistent in their choice of document classes and when they reformulated and tried out alternative facets and concepts. Altogether the findings demonstrate that doctors were able to operate using the SC model, and that the model was a helpful feature to structure and express queries specified to domain-related aspects. SCs were applied in all high performance queries except one, demonstrating the model's usefulness.

Concluding remarks

Workplace queries are typically context-specific and often limited to domain-specific aspects of the search topic. The goal of the study reported here was to gain insight into how doctors applied the SC model to formulate queries to solve work-related search tasks. The results showed that doctors used the model purposively and interactively when choosing search facets and search concepts. As intended they used the model to specify the search within context-specific aspects. They were relatively consistent in their use. Compared to the baseline system the SC model produced the highest performance scores over all tasks. The findings provide promising evidence of the potential usefulness. In future research we would like to investigate the SC model in other domains and settings, especially to find out whether users that are not necessarily domain experts can benefit from the highly specified model.

References

Aitchison, J., Gilchrist, A. & Bawden, D. (2002). *Thesaurus construction and use*. London: Fitzroy Dearborn.

Bishop, A. Document structure and digital libraries: How researchers mobilize information in journal articles. *Information Processing and Management*, 35, 225-279.

Crowston, K. & Kwasnik, B. H. (2003). Can document-genre metadata improve information access to large digital collections? *Library Trends,* 52(2), 345-361.

Dillon, A. (1991). Reader's models of text structures: the case of academic articles. *International Journal of Man-Machine Studies*, 35, 913-925.

Ely, J.W., et al. (1999). Analysis of questions asked by family doctors regarding patient care. *British Medical Journal*, 319, August. 358-361.

Fagin, R., et al. (2003). Searching the workplace web. In: *Proceedings of the 12th International World Wide Web Conference (WWW '03),* Budapest, Hungary, May 20-24, 2003. 366-375.

Freund, L., Toms, E. & Waterhouse, J. (2005). Modeling the information behaviour of software engineers using a work-task framework. In: Grove, A (ed.) *ASIS&T '05 Proceedings of the 68th Annual meeting,* Charlotte, NC, October 28-December 2, 2005.

Price, S. L., et al. (2009). Using semantic components to search for domain-specific documents: an evaluation from the system perspective and the user perspective. *Information Systems,* 34 (8), 778 – 806.

Turner, A.M., Liddy E.D., Bradley, J., & Wheatley, J.A. (2005). Modelling public health interventions for improved access to grey literature. *Journal of the Medical Library Association*, 93 (4), 487-494.

Vickery, B. C. (1960). Faceted classification: a guide to construction and use of specialized schemes. London: Aslib. 1968.

Corresponding author:

Marianne Lykke can be contacted at mlykke@hum.aau.dk