Moving Toward Web-Scale: Adapting Semantic Components for Use in Large Collections

Susan Price
Department of Computer Science
P.O. Box 751
Portland State University
Portland, OR 97207, USA
prices@cs.pdx.edu

Abstract: Some users' information needs are very targeted, especially in domain-specific settings. The semantic components model supplements traditional full text and keyword indexing with a semantic description of subdocument content that does not necessarily correspond to structural elements in a document. The model extends typical query languages by allowing user queries to address subdocument components in addition to the whole document. We have evidence from a large interactive user study showing that semantic components can enhance document retrieval in a domain-specific digital library. We now propose to adapt the semantic components approach to improve its scalability for use in large document collections by allowing user indexing, by allowing multiple indexing instances per document, and by introducing an open semantic component schema. The proposed research will examine design issues and implementation options and provide preliminary evaluations of the effect of these adaptations on retrieval performance.

Keywords: Information retrieval, indexing, digital libraries, user studies, semantic components, collaborative indexing.

1. SEMANTIC COMPONENTS: AN INTRODUCTION

Some users' information needs are very targeted. A busy user may want a single document that contains specific information needed to answer a question or support a decision. Often the desired information is about a particular aspect of a larger topic. This is particularly true when domain experts already have a large fund of knowledge about the domain and need information about a particular aspect of a familiar topic. Consider, for example, a physician who suspects her patient has a pheochromocytoma, a rare tumor, but does not remember the diagnostic criteria. She does not want an overview because she already knows what a pheochromocytoma is. She wants targeted information to help her decide what tests to order today. Domain experts also have knowledge about what kinds of documents exist, how documents of various types are organized, and where they can be found.

We recently introduced the *semantic components* model (Price et al., 2006a, Price et al., 2006b), which seeks to leverage searchers' existing knowledge about a domain by providing an enhanced document representation and extended query language that supplements existing full text (and keyword, if present) indexing. The model has two main elements: document classes and semantic components. Different domains and document collections may have different axes that are most appropriate for classifying documents. In health-related collections we have found topic type to be very useful. For example, such collections often have documents about diseases (one class) and documents about drugs (another class). Documents within a class tend to contain characteristic types of information. For example, documents about diseases often contain information about *treatment* and about *diagnosis*. Documents about drugs generally contain information about *dosage* and about *interactions* with other drugs. We call these types of information *semantic components*. We call the set of document classes and associated sets of semantic components that are identified for a particular document collection a *semantic component schema*. A *semantic component instance* is the text in a document that contains information about the subtopic that is the semantic component. Semantic component instances may or may not correspond to structural elements in documents, can overlap with other instances, and may consist of discontinuous segments of text. Any given text in a document can belong to zero, one, or many semantic component instances.

We use semantic components in three ways:

1. We allow searchers to search for query terms in specific semantic components in addition to searching for topical query terms in whole documents. A query could consist of the topical term "pheochromocytoma" and the term "criteria" applied to a *diagnosis* semantic component.
2. We allow searchers to specify a preference for documents containing particular semantic components, without searching for a particular term within a semantic component. In this case, a query could consist of the topical term "pheochromocytoma" and a request for documents containing an instance of the *diagnosis* semantic component.

3. We display information about the presence and size of semantic components in documents that appear in search results. One document about pheochromocytoma might contain an instance of *treatment* consisting of 100 words and an instance of *diagnosis* consisting of 200 words while another document might contain only an instance of *treatment* with 500 words. One can also imagine returning to the user only the matching semantic component instances from a search, or returning whole documents but with the relevant semantic component instances highlighted for easy identification.

We have evidence indicating that semantic components are useful for retrieving documents from a domain-specific digital library (Price et al., 2007). We recently completed a large interactive searching study in which 30 domain experts searched a collection of nearly 25,000 documents for four realistic search scenarios. We compared a basic search to an experimental system with a prototype implementation of semantic components. The basic system used existing full text and keyword indexing and mimicked an operational system familiar to the searchers. The experimental system used the same full text and keyword indexing and, in addition, implemented semantic components on top of the basic system. The participants used each system for two scenarios. Searchers issued queries with better search performance using semantic components than when using the basic system, as measured by normalized discounted cumulative gain (nDCG) and mean average precision (MAP) using a reference standard based on expert relevance judgments. We are also analyzing data from the user perspective, considering the searchers' own relevance assessments and qualitative feedback. Other recent work has: (1) demonstrated the feasibility of using the semantic components framework by developing semantic component schemas for three collections in two different domains (Price et al., 2006a, Price et al., 2006b), and (2) demonstrated that information needs can be expressed using semantic components by mapping an existing taxonomy of information needs of family practice physicians to the semantic component schemas developed for two appropriate document collections (Price et al., 2006a). For semantic components to be useful, classifying documents and identifying the presence and location of semantic components in documents must be accurate, consistent, and scalable. We recently completed a study that compared manual semantic component indexing to manual keyword indexing. Data analysis is still in progress, but preliminary results suggest that the two types of indexing are similar with respect to time and intellectual effort.

We propose to adapt the semantic component approach for use in larger collections, including the Web. We have considered two main approaches to achieving semantic component indexing in large document collections: (1) automated indexing, using tools such as machine learning and natural language processing, and (2) harnessing the attention of document users. While both techniques may prove useful, in this paper we only consider the latter approach. Our vision is to adapt semantic components for use in arbitrarily large document collections, including the Web. However in this project we intend to study scalability issues and test scalable solutions in the context of smaller, more well-defined document collections in order to control experimental variables.

2. SEMANTIC COMPONENT INDEXING

We use the term *semantic component indexing* for the process of associating segments of text in a document (i.e., semantic component instances) with semantic component names, or labels. We refer to the stored data about semantic component instances as the *index*, with no explicit requirements about how such storage is implemented.

2.1 Document indexing

For the recently completed searching study, we used a prototype indexing tool that allows a human indexer to highlight segments of text, to right-click in order to display and select from a menu of semantic component names, and to see the results of indexing in progress. Additional segments of text can be added to an existing semantic component instance by repeating the highlight and right-click procedure. Figure 1 shows a screenshot of the indexing tool after the document (in Danish) has been classified and some segments of text have been associated with semantic components. Text is automatically highlighted in the color associated with the chosen semantic component and the text is copied into the appropriate pane on the right to provide visual cues regarding what indexing has already been assigned.

2.2 Harnessing the attention of document users

Traditionally, indexing has been the job of professional indexers. Indexing is difficult to do well and even professional indexers are inconsistent when applying keywords from controlled vocabularies (Funk and Reid, 1983). More recently the phenomenon of collaborative tagging has emerged on the Web (Macgregor and McCulloch, 2006). Creators and users can assign descriptors, commonly referred to as tags, to describe a variety of electronic resources such as web pages¹, bibliographic references², and photo collections³. We propose to harness the intellectual efforts of users in a similar way, extending our model to allow user indexing. Our proposal is based on three observations: (1) The phenomenon of collaborative tagging suggests that (some) users will volunteer time and effort to categorize resources. (2) Our comparison of manual keyword indexing and semantic component indexing suggests that a large proportion of the time and effort of indexing is attributable to reading and comprehending the document. Presumably the volunteer user/indexer has already committed time to this step. We conjecture that selecting and labelling semantic component instances will take relatively little additional effort.

¹ <http://del.icio.us>

² <http://www.citeulike.org/>

³ <http://flickr.com>

(3) During a usability study of Metadata++ (Weaver et al., 2007, forthcoming), a digital library system with a path-based thesaurus for indexing, searching, and browsing, domain experts exhibited a striking familiarity with the organization of long documents, rapidly homing in on critical sections of interest (Delcambre, 2007). This, in conjunction with the work of Dillon (Dillon, 1991) and Bishop (Bishop, 1999) demonstrating users' ability to manipulate subdocument components, suggests that users who are familiar with a domain and its document types may excel at selecting and classifying subdocuments that will be useful to other users.

2.3 Allowing multiple instances of semantic component indexing

Allowing multiple instances of semantic component indexing for a single document serves two purposes: (1) it allows representation of document content from multiple user perspectives, and (2) it may improve indexing quality. Although the semantic component model does not require that a document belong to only one document class or have only one instance of indexing, we have so far implemented the model with at most a single indexing instance per document. In our experience, document classification is not perfect. Not every document belongs obviously to a single class, and indexers do not always agree on the single most appropriate class for a document. Indexers also do not always select the same text segments, with the same boundaries, when indexing semantic components. While some disagreement can be expected in any task requiring human judgment, there may also be multiple appropriate semantic component indexing instances for a given document. The same document might be useful for multiple purposes, by multiple user groups. Different target audiences might have different perspectives and be better served by different semantic components.

At first glance, delegating indexing to users may seem risky because user indexing is unpredictable and uncontrolled. Yet while one instance of indexing may be unreliable, the accumulation of multiple indexing instances is likely to converge toward a meaningful result and may, on average, be better than indexing produced by a single individual. Studies of del.icio.us provide evidence that tagging by a critical mass of users results in convergence to stable tag usage patterns that can be described by a power law distribution, possibly due to a combination of imitation and shared knowledge (Golder and Huberman, 2006, Halpin et al., 2007). Furthermore, semantic components are supplementary to traditional indexing and search, allowing the user to more precisely specify a search. Poor user indexing will inhibit the ability of semantic components to improve search precision, but is unlikely to significantly degrade retrieval quality compared to traditional whole document search alone.

Whereas differences among instances of collaborative tagging relate only to tag selection, disparities among semantic component indexing instances will relate to boundaries of semantic component instances as well as labels. An important area of research is determining how to best combine multiple indexing instances over the same document for retrieval purposes. We may want to consider how many indexing instances contribute to defining each semantic component.

2.4 Allowing multiple semantic component schemas

Until now we have been investigating semantic components in the context of domain-specific digital libraries. We create a semantic component schema by identifying document classes and semantic components that describe a particular document collection and that may be useful for information retrieval. Developing the schema requires time and intellectual effort that must be repeated for each document collection. Furthermore, the schema must be created in advance, and once created it is fixed. The current model has no provision for a semantic component schema to evolve as the nature of a collection changes or the state of knowledge within a domain changes. Similar to keyword indexing, successful use of semantic component indexing requires a shared understanding of document class names and semantic component names between indexers and searchers. This shared understanding is less likely in larger collections that span a broader range of subjects. We propose to extend the reach of the semantic component approach by modifying our model to eliminate the "schema first" requirement, providing increased flexibility and eliminating the need for a pre-existing shared understanding of labels.

In order to scale up to large and unrestricted document collections, such as the Web, we introduce the notion of an *open* semantic component schema. When a semantic component schema is defined in advance and remains fixed, we call this a *closed* semantic component schema. An open semantic component schema has two

Figure 1: Semantic component indexing application.

characteristics: (1) An open schema has only one level, semantic components. We discard the two-level hierarchy. (2) An open schema has no predefined semantic components. An indexer may associate a segment of text with any name deemed appropriate. The open schema approach resembles collaborative tagging except that a tag is bound to a whole document whereas a semantic component name is bound to a selected subdocument. The open schema retains the essence of the semantic component approach, which extends whole document search by also searching subdocuments, where subdocuments are defined on a semantic, not structural, basis.

3. RESEARCH QUESTIONS

Our vision is to explore application of the semantic components approach to arbitrarily large document collections. Here we propose steps toward achieving the broader goal, research designed to answer the following questions:

1. What are the challenges, design trade-offs, and implementation options for allowing multiple semantic component indexing instances? This question relates to both (a) efficiently storing multiple semantic component indexing instances and (b) incorporating multiple semantic component indexing instances into algorithms for retrieving and ranking documents in response to a query. Initially we will explore issues in the context of a closed semantic component schema.
2. What are the challenges, design trade-offs, and implementation options for allowing an open semantic component schema? This question also relates to both efficient data storage and effective algorithms for retrieval and ranking.
3. When compared to a single indexing instance per document, produced by an experienced indexer, what effect does allowing user indexing and multiple indexing instances per document have on (a) the time and perceived difficulty of semantic component indexing and (b) the retrieval performance of a search system?
4. When compared to a fixed semantic component schema, what effect does an open semantic component schema have on (a) the time and perceived difficulty of semantic component indexing and (b) the retrieval performance of a search system?

In addition, we have an associated question and goal related to evaluation methodology that is both essential to evaluating our work and also potentially useful to other researchers investigating new forms of indexing.

5. How can we assess the retrieval performance of a search system using semantic components in the absence of complete indexing? By complete indexing, we mean at least one semantic component indexing instance per document.

4. PROPOSED RESEARCH METHODS

4.1 Exploring design alternatives for allowing multiple indexing instances with a closed schema

Previously we implemented semantic components on top of a commercial search engine that was being used for the operational system that we wanted to mimic for our searching study. We implemented searching with semantic components as document re-ranking based on a secondary search of metadata fields in which we stored semantic component indexing. For this project we will systematically consider implementation alternatives using an open source search engine API, such as Lucene. In particular we want to investigate the complexities introduced by allowing multiple indexing instances that may or may not include assignment to the same document class. How can we efficiently store multiple indexing instances per document? How does the system scale, with respect to time (to compute document matching) and space (to store an index), as the average number of instances per document increases? How can multiple indexing instances be combined for document retrieval and ranking?

4.2 Exploring design alternatives for allowing semantic component indexing with an open schema

This part of the work will build on the research exploring alternatives for allowing multiple indexing instances with a closed schema. We will answer the same questions about the effect of an open schema on scaling, with respect to time and space. Combining instances using an open schema for the purpose of retrieval will be even more challenging than for a closed schema. Should semantic component names be considered as entirely independent of each other? Should semantic component instances be combined if they are labelled with identifiable synonyms? Should hierarchically related terms be combined? Should other linguistic tools, such as a stemmer or lemmatizer be used? We may also explore statistical methods for clustering semantic component labels.

4.3 User experiment to assess the effect of allowing multiple indexing instances with a closed schema

We will explore the effects of allowing multiple indexing instances with a closed schema in the context of our recent searching study. We will recruit multiple users to index documents that we used in that study, measuring the time required to index each document, and using a questionnaire to assess the perceived difficulty of indexing. We will then re-use the queries and relevance judgments from the searching study to assess alternatives for combining indexing instances and to obtain preliminary evidence about whether allowing multiple indexing instances enhances or degrades retrieval performance.

4.4 User experiment to assess the effect of allowing semantic component indexing with an open schema

We will initially use the searching study context to explore open schemas by recruiting users to index with an open schema the same documents already indexed with a closed schema. This will allow us to study consistency with respect to choice of semantic component names and overlap of segments of text. We can combine this user experiment with the one described in Section 4.3 by asking each participant to index half of the documents with an open schema and half with the existing closed schema. We will assign the open schema indexing task first so that choice of labels will not be influenced by exposure to the closed schema.

Because the queries entered by the participants in the original searching study are specific to the existing semantic component schema for the document collection, we cannot re-use the queries against documents indexed with an open schema. Instead, we will additionally ask participants to generate one to three queries for each of the four original search scenarios. This will allow us to asynchronously study real user queries for realistic scenarios before progressing to a more expensive interactive user study.

4.5 Comparing the effect of fixed and open semantic component schemas

After preliminary research has established good implementation options for allowing multiple indexing instances and open schemas, we will compare the effect of closed and open schemas on retrieval. For this experiment we will investigate semantic components in a new document collection and a new domain (yet to be identified). This comparison will involve a preparatory phase and two experiments. The preparatory phase will consist of developing a semantic component schema and realistic search scenarios. The first experiment involves indexing. We will ask participants to index documents using both a closed and an open schema. We will study consistency, time, and perceived effort required for indexing. The second experiment will be an interactive searching study, using the indexing instances from the indexing experiment and additional indexing instances as needed to form a critical mass of indexed pages. The participants will search half of the scenarios using the closed schema and half using the open schema so that we can compare the retrieval results for the two schema types.

4.6 Evaluation in the presence of partial indexing

Manual indexing is expensive. Complete manual indexing of a large document collection for experimental use is not feasible, making it challenging to evaluate any new approach to indexing. Even if the ultimate intent is to automate indexing, it can be useful to first evaluate the usefulness of a humanly produced version to ensure quality. For our recently completed searching study we prospectively identified as many relevant documents as possible for each scenario and also identified the set of documents (not relevant) most likely to be returned by searches for each experimental scenario. We indexed the relevant documents plus this set of documents that we expected would most likely compete for ranking with relevant documents. After the study, we retrospectively analyzed the results of our selective indexing to ensure that the selection of documents to be indexed did not bias the results of the study. Our analysis depended on two characteristics of the experiment: (1) there were only a few relevant documents for each scenario, and (2) the experimental system was re-ranking documents returned by the baseline comparison system. We plan a more general analysis to formulate a method for determining whether document selection for indexing has biased experimental results when queries are not known in advance.

5. RELATED WORK

Document classes are related to document genre, which other authors have suggested may improve information retrieval (such as (Crowston and Kwasnik, 2003, Rauber and Müller-Kögler, 2001, Freund et al., 2005)). Semantic components are similar to facets used for classification and to indexing keywords (which may represent facets). Medical literature is often indexed with the Medical Subjects Heading (MeSH) vocabulary,⁴ which contains both descriptors (the main terms) and qualifiers. When used together, a descriptor/qualifier pair represents a particular aspect of a topic, much like a semantic component. For example, the descriptor/qualifier pair *appendicitis/etiology* might be used to index a document about appendicitis (a disease) that discussed etiology (a semantic component for the class of documents about diseases). An important difference is that keywords are bound to the document as a whole whereas semantic component instances are bound to particular segment(s) of text in the document.

Many domain-specific documents are written in a highly characteristic fashion. Purcell et al. developed context models to represent the types of information in medical research articles and used them to represent documents in a retrieval system (Purcell et al., 1997). Their context-based representation is similar to the semantic components model, but is more closely tied to document organization than our semantic approach. When document structure corresponds to semantic content, structure can be used to efficiently identify semantic components. If structure is explicit, as with XML markup, structure can be exploited directly. With an existing closed schema, new documents could be authored such that structural elements directly represent semantic components. With an open schema, XML structural elements could be used directly as well, although their usefulness might be unpredictable.

A variety of IR-related tasks involve subdocument granularity. Those most closely related to semantic components are content analysis (Krippendorff, 2004), text segmentation used to aid information retrieval (Hearst and Plaunt, 1993) and to display retrieval results (Hearst, 1997), and passage retrieval. Semantic component instances might be considered as semantic passages (Liu and Croft, 2002), although not all document text is necessarily part of a semantic component instance. Unlike passage retrieval, we use information about semantic component instances to supplement, not replace, whole-document retrieval techniques.

A number of authors have studied collaborative tagging and folksonomies. For example, Mika has studied the emergence of lightweight ontologies that can be extracted from del.icio.us (Mika, 2005), and Hotho et al. have proposed FolkRank to exploit folksonomies for search and ranking (Hotho et al., 2006).

⁴ <http://www.nlm.nih.gov/mesh/>

6. ISSUES FOR DISCUSSION

1. Are there existing test collections that can be (partially) leveraged for studying semantic components before progressing to full-scale user studies?
2. Can existing queries that relate to relationships between concepts be leveraged for studying retrieval performance, converting them into queries against an open semantic component schema?
3. Are there existing techniques that would be useful for studying retrieval performance in the absence of complete indexing?

7. ACKNOWLEDGMENTS

The work on semantic components described here is joint work with my advisor, Lois Delcambre, at Portland State University, and Marianne Lykke Nielsen at the Royal School of Library and Information Science in Denmark. The recent searching and indexing studies described were done with the support and collaboration of sundhed.dk, the national Danish health portal (<http://www.sundhed.dk>). The semantic components work was supported in part by the National Science Foundation, grant numbers 0514238, 0511050, and 0534762 and by the National Library of Medicine Training Grant 5-T15-LM07088. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES.

- [1] Bishop, A. (1999) Document structure and digital libraries: How researchers mobilize information in journal articles. *Information Processing & Management*, **35**, 225-279.
- [2] Crowston, K. and Kwasnik, B. H. (2003) Can document-genre metadata improve information access to large digital collections? *Library Trends*, **52**.
- [3] Delcambre, L. (2007) Personal communication.
- [4] Dillon, A. (1991) Reader's models of text structures: the case of academic articles. *International Journal of Man-Machine Studies*, **35**.
- [5] Freund, L., Toms, E. G. and Clarke, C. L. (2005) Modeling task-genre relationships for IR in the workplace. *Proceedings of SIGIR '05*, Salvador, Brazil, August 2005.
- [6] Funk, M. E. and Reid, C. A. (1983) Indexing Consistency in MEDLINE. *Bulletin of the Medical Library Association*, **71**, 176-183.
- [7] Golder, S. A. and Huberman, B. A. (2006) The structure of collaborative tagging systems. *Journal of Information Science*, **32**, 198-208.
- [8] Halpin, H., Robu, V. and Shepherd, H. (2007) The complex dynamics of collaborative tagging. *WWW 2007*, Banff, Alberta, Canada, May 8-12, 2007.
- [9] Hearst, M. and Plaunt, C. (1993) Subtopic structuring for full length document access. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA, 1993.
- [10] Hearst, M. A. (1997) TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, **23**, 33-64.
- [11] Hotho, A., Jäschke, R., Schmitz, C. and Stumme, G. (2006) Information retrieval in folksonomies: Search and ranking. *European Semantic Web Conference (ESWC 2006)*, Lecture Notes in Computer Science 4011: Springer-Verlag, pp. 411-426.
- [12] Krippendorff, K. (2004) *Content analysis: An introduction to its methodology*, Sage Publications, Thousand Oaks, CA.
- [13] Liu, X. and Croft, W. B. (2002) Passage Retrieval Based On Language Models. *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02)*, McLean, VA, November 4-9, 2002.
- [14] Macgregor, G. and McCulloch, E. (2006) Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, **55**, 291-300.
- [15] Mika, P. (2005) Ontologies are us: A unified model of social networks and semantics. *International Semantic Web Conference (ISWC 2005)*, Lecture Notes in Computer Science 3729: Springer-Verlag, pp. 522-536.
- [16] Price, S. L., Delcambre, L. M. and Nielsen, M. L. (2006a) Using semantic components to express clinical questions against document collections. *International Workshop on Healthcare Information and Knowledge Management (HIKM 2006)*, Arlington, VA, November 11, 2006.
- [17] Price, S. L., Delcambre, L. M., Nielsen, M. L., Tolle, T., Luk, V. and Weaver, M. (2006b) Using Semantic Components to Facilitate Access to Domain-Specific Documents in Government Settings. *The 7th Annual International Conference on Digital Government Research (dg.o)*, San Diego, California, May 21-24, 2006.
- [18] Price, S. L., Nielsen, M. L., Delcambre, L. M. and Vedsted, P. (2007) Semantic components enhance retrieval of domain-specific documents. *Conference on Information and Knowledge Management (CIKM 2007)*, Lisbon, Portugal, November 6-9, 2007 (accepted).
- [19] Purcell, G. P., Rennels, G. D. and Shortliffe, E. H. (1997) Development and evaluation of a context-based document representation for searching the medical literature. *International Journal on Digital Libraries*, **1**, 288-296.
- [20] Rauber, A. and Müller-Kögler, A. (2001) Integrating automatic genre analysis into digital libraries. *Proceedings of the 1st ACM/IEEE-CS joint conference on digital libraries (JCDL '01)*, Roanoke, VA, 2001.
- [21] Weaver, M., Delcambre, L. M., Nielsen, M. L., Price, S. L., Maier, D. and Tolle, T. (2007, forthcoming) Supporting domain-specific digital libraries in government: Two case studies. In Chen, H., et al. (eds), *Digital Government: Advanced Research and Case Studies*, Springer, New York.