

Automatically Associating Documents with Concept Map Knowledge Models

Alejandro Valerio and David Leake

Computer Science Department, Indiana University
Lindley Hall 215, Bloomington, IN 47405, USA
{avalerio,leake}@cs.indiana.edu

and

Alberto J. Cañas

The Institute for Human and Machine Cognition
40 S. Alcaniz St, Pensacola, FL 32502, USA
acanas@ihmc.us

Abstract

Concept-map-based knowledge models are widely used for knowledge capture and sharing. When libraries of concept maps are available, those concept maps can provide a useful context for understanding new documents, and the new documents can provide useful annotations to the knowledge models—if the right documents can be associated to the right concept maps. This paper presents ongoing research on a method for classifying documents according the most relevant concept maps in a concept map library, using natural language processing techniques. It presents an algorithm for extracting concept map fragments from a document, introduces a matching procedure, and compares performance of the overall algorithm with a baseline method, with encouraging results.

Keywords: Concept mapping, document classification, knowledge representation, natural language processing

1 Introduction

The explosive growth of on-line document repositories provides strong motivation for developing tools to help people to select and understand relevant documents. This paper describes ongoing research on methods for extracting concepts from documents, in order to use the extracted information to classify documents into categories defined by human-generated *concept maps* [22].

Concept mapping has been widely used in education and knowledge capture, as a method for formulating rich structured descriptions of knowledge for knowledge examination, sharing and browsing (see [24, 4] for a sampling of this work). Concept maps express concepts and relationships in a two-dimensional network, described in informal terms rather than in a formal representation: They use natural language for concept and link labels, and the concept-link-concept triples of concept maps form simple natural language sentences. Part of the appeal of concept maps is that domain experts can generate concept maps without the intervention of a knowledge engineer, providing a direct source of descriptions of experts' conceptualizations.

Electronic representations of concept maps may have a variety of attached resources including images and documents, constituting a rich information environment to assist users understand the document in context. Methods to automatically form associations between documents and electronic concept maps are desirable to help concept-map-builders to enrich electronic concept maps by annotating them with relevant documents. Likewise, they could help the users of concept maps by supporting retrieval of new documents relevant to the content of a concept map being consulted. Conversely, when a user is consulting a document, the ability to associate concept maps with documents could be used to retrieve concept maps relevant to a document of

interest. Our current work addresses the fundamental problem for providing the needed functionality: the classification problem in which concept maps are the classes and the task is to assign the document to the most relevant of a set of concept maps.

Successful classification depends on natural language processing for both candidate documents and concept maps. Our goal is to develop a domain-independent approach, which makes the task more difficult by precluding assuming deep knowledge or the availability of the a priori information on which information extraction systems sometimes rely (e.g., [17]). However, because the goal is classification, it is not necessary for the generated information to be provide a perfect representation of the content of the text: It suffices to achieve sufficient accuracy to associate the document to the more relevant concept map in a set of candidates.

This paper begins by briefly summarizing concept maps and the use of electronic concept maps as a vehicle for knowledge construction and sharing. It then presents our algorithm, which builds on our work to construct concept maps from documents by applying in an updated, simplified version of [29]. It then presents an evaluation comparing the new algorithm to a baseline, with encouraging results. Finally it discusses the context for our work with a summary of relationships to prior work on concept mapping related tasks, document classification and NLP.

2 Concept maps as a tool for knowledge construction and sharing

As illustrated in Figure 1, concept maps represent the concepts and relationships of a domain in a two-dimensional network. Nodes in the network correspond to concepts, and links correspond to concept relationships. Concept maps were developed in the context of education [22], with the process of building concept maps seen as a way to facilitate human knowledge construction, organization, and sharing. As students generate material to include in concept maps, they construct and refine their own understanding; by expressing their understanding in an explicit visual form, they make it available for others to examine and compare.

More recently, concept mapping has been recognized as a useful tool for knowledge construction and sharing by domain experts. In contrast to formal network knowledge representation models explored in artificial intelligence, such as semantic networks, conceptual graphs, and text graphs [23], concept maps are informal; both concept and link labels are expressed as simple sentences in natural language. This facilitates the knowledge capture process needed to build concept maps, enabling experts to capture their knowledge directly or with only limited assistance from a knowledge engineer.

The ability of experts to construct concept maps directly has resulted in many projects to capture expert knowledge in the form of concept-map-based *knowledge models*, collections of topically related linked concept maps with rich annotations such as documents or images (e.g., [5]). The CmapTools concept mapping software [9] facilitates this process by providing a vehicle for generation and sharing of electronic concept maps annotated with additional resources. The CmapTools interface and a sample concept map relating machine learning, natural language processing and information retrieval are shown in Figure 1.

Despite the relative ease of concept map generation, determining the right content for concept maps may still be difficult. This has motivated efforts to develop methods to support the concept mapping process (e.g., [19]). In particular, when annotating the concept maps in a knowledge model with documents from a large document set, identifying relevant documents to associate with a given concept map may be prohibitively time-consuming. The work described in this paper is aimed at addressing that problem by automating the process of associating documents to concepts.

3 Automatic extraction of document information for matching with concept maps

Many natural language processing techniques exist for exploiting the information contained on the structure of sentences and phrases on documents [14, 6, 2]. For example, question answering and information extraction systems require the identification of entities and their relationships. Part of our research examines how to adapt such existing algorithms to the specific needs of automatically constructing concept maps from documents, as described in [29]. That paper presented initial work on methods whose goal was to generate concept maps resembling those authored by humans.

For our current task of associating documents to existing concept maps, many of the same methods are relevant, and the current algorithm is a modification of our previous algorithm. However, unlike that work,

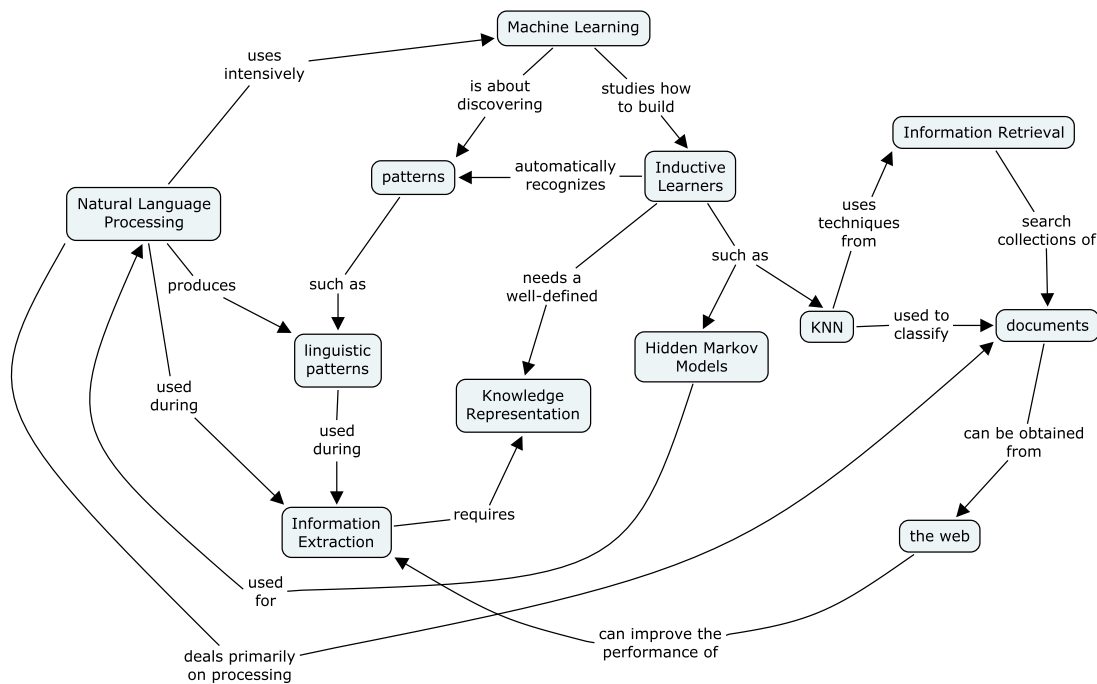


Figure 1: Sample cmap displayed by CmapTools

in which a concept map generated from a document was the intended product, here the concept map is generated for internal system use, as an index. Thus the concept map is an alternative representation of the document’s content, in which the concept connectivity is used to estimate the relevance of concepts in the document for term weighting.

Because the current task is to identify the most relevant of a set of existing concept maps, the current model follows a more top-down approach, removing the need for bottom-up concept weighting and selection based on the document alone. Instead, all potentially relevant concepts in the document are retained.

The current algorithm also uses a partial syntactic parser trained to find sentence segments corresponding to concepts and linking phrases, substituting the functions of the deep syntactic parser. This provides significant improvements in the algorithm execution time.

The algorithm used for this task is summarized on Figure 2. The algorithm steps are described below.

Parsing: The document is first preprocessed by an ad hoc sentence boundary detection algorithm based on regular expressions, followed by a part-of-speech tagger that annotates each word according to its syntactic role on the sentences (as nouns, verbs, adjectives, ...).

Each sentence is then processed by a partial parser to recognize sequences of words corresponding to concepts and linking phrases, using the part-of-speech tags as input. The parser uses a modification of Abney’s partial parser [1], replacing the chunk-level grammar in that work with one that recognizes the syntactic structure of concept map propositions. This new grammar was produced by automatically analyzing a set of propositions from human constructed concept maps, determining the most likely part-of-speech sequences on concepts and linking phrases. This analysis is made based on the frequency of these sequences in the propositions. This tool was used as a faster alternative to the previous full parsing procedure [11], which gives a detailed syntactic analysis of the sentence.

Word normalization: Documents contain morphological variations of words that refer to the same entity, and may use multiple synonyms. The normalization step splits words into disjoint equivalence classes, in which two words a and b are considered equivalent if $POS(a) = POS(b)$, and $lemma(a) = lemma(b)$ or $lemma(a) \text{ synonym } lemma(b)$. $POS(a)$ is the part-of-speech of a and $lemma(a)$ is the root word of a (e.g., the root word of “realizing” is “realize”). WordNet [13] is used to determine the synonymy relation. Occasional conflicts may arise because words may have more than one meaning, in which case the algorithm

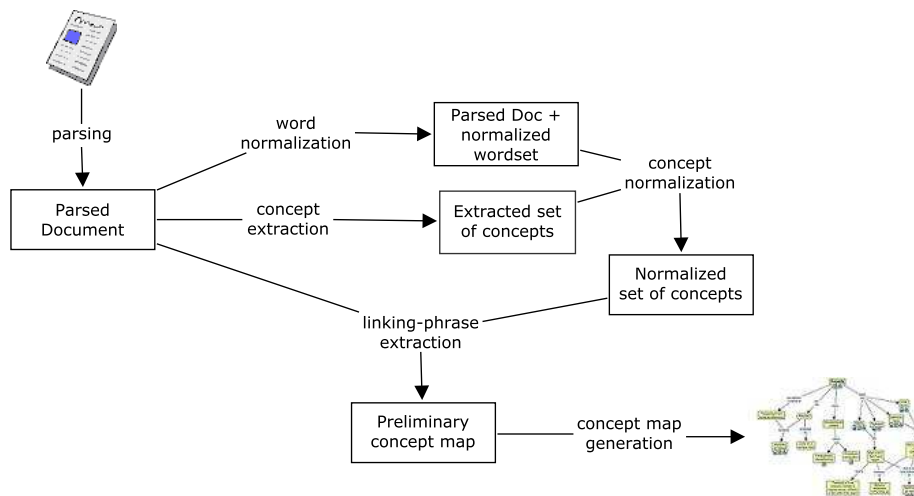


Figure 2: Procedure to construct a concept map automatically from a document

arbitrarily assigns the word to one of the possible classes. Once the algorithm identifies the word equivalences, it tags each word with its class, for use to compare words in later steps.

Concept extraction: This step simply selects the concepts discovered during the partial parsing.

Concept normalization: The sentence chunks corresponding to concept labels may have superficial differences although some of them may refer to the same concept. This step implements a simple solution to filter these differences and integrate equivalent concepts. Two concept labels are considered the same if all nouns and adjectives in one are contained in the other, considering the classes produced during the word normalization step.

Linking phrase extraction: Using the parsed sentences and normalized concepts, the sentences are searched for linking phrases that appear between two concepts. These three chunks are used to generate a proposition, as we presume that the phrases show relations between concepts.

Concept map generation: The final step gathers the information from the extracted concepts and linking phrases in the form of propositions to construct a graphical representation of the concept map. Although the graphical representation is not required to construct the concept map index from the document, it enables the results to be displayed by existing tools for concept map construction and refinement such as CmapTools from The Institute for Human and Machine Cognition [9]. The CmapTools Knowledge Exchange Architecture (KEA) [8] allows a seamless integration of external tools to the CmapTools knowledge base and has auto-layout capabilities, facilitating displaying the resulting concept map.

Figure 3 illustrates the output of processing a document on the subject of “light bulbs”.

4 Matching an input document with the closest concept map

The matching phase examines a collection of concept maps, constructed manually by human experts, and find the map that is most closely related to the topic of an input document. In this section we describe the process in detail. First, the system applies the algorithm described on Section 3 to produce a concept map from the document. This concept map becomes the document index.

To identify relevant concept maps, the document concept map is compared with all maps in the target collection using cosine similarity [3] and a vector-model representation of concept maps [19]. The concept map vectors are constructed as follows. Using the Hub-Authority-Root-Distance (HARD) [27] model for estimating concept importance based on structural features, each concept is assigned a weight based on its

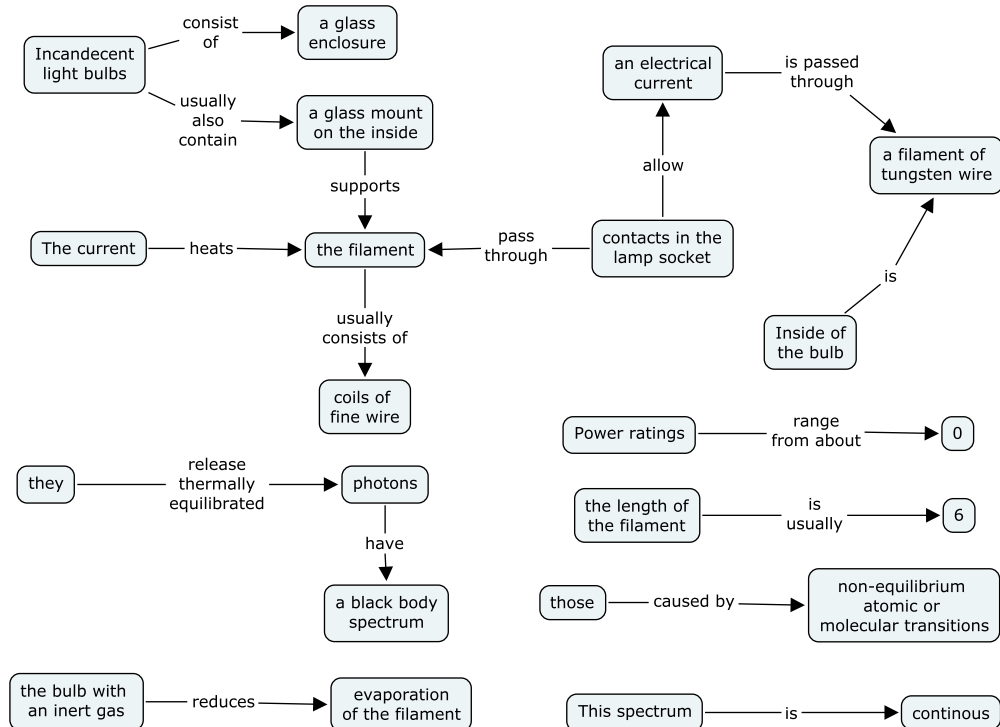


Figure 3: Output from processing a sample document concerning light bulbs.

authority value (increasing with number of incoming connections from hubs), *hub* value (increasing with number of outgoing connections to authorities), and *upper node* value (shortest distance to root concept). The HARD model has been shown to be a good model of the importance humans assign to concepts in a map based solely on its layout [27]. Next, individual keywords are assigned weights according to their frequency and the weight of concepts in which they appear. Each keyword defines a dimension in the concept map vector.

More specifically, the weight $w(i)$ of concept i according to the HARD model is:

$$w(i) = \phi \cdot h(i) + \psi \cdot a(i) + \gamma \cdot u(i)$$

where $h(i)$, $a(i)$, and $u(i)$ are the authority, hub, and upper node values for i , described in detail in [10]. The parameters have been set to $\phi = 0$, $\psi = 2.235$, and $\gamma = 1.764$, which were previously found to best fit the model for experimental user data [18]. The weight $w(j)$ of keyword j is the sum of the concept weights multiplied by the frequency of the keyword in each concept.

$$w(j) = \sum_{i \in \text{concepts}} \text{frequency}(i, j) \cdot w(i)$$

Keywords are normalized with a lemmatizer to prevent mismatches due to morphological variations and also tagged with part-of-speech to reduce noise.

5 Experimental setup and evaluation

Our evaluation focuses on determining the ability of the algorithm to associate an input document to the most relevant map in a collection of concept maps constructed by experts. To test the performance of the procedure, we used existing knowledge models containing a large number of concept maps annotated with topically related documents. In our test, all documents are first removed from these concept maps. Next, each of the documents is processed individually, with no prior knowledge about the concept map to which it was originally linked. The evaluation is based on a match between the concept map identified by the tool as

the most relevant and the original concept map annotation, measuring the ability of the procedure to find the original association.

An implicit assumption is made about the association of documents to concept maps in the expert knowledge models used for evaluation: a concept map is linked to all documents that are relevant to it, making it possible for a document to be associated with more than one concept map. This assumption may not always hold, because the data may have errors and noise, but those are conditions generally found in well constructed knowledge models.

We tested the system using 54 concept maps and 63 attached documents from the Mars 2001 knowledge model [5], and 26 concept maps with 78 documents from the STORM-LK knowledge model [15], which have been used previously as a “gold standard” expert-generated concept maps. The tool processed each document by reattaching it to one of the 80 concepts maps in the collection according to the similarity measure described above. We consider an attachment to be successful if the document is correctly associated with any of the concept maps originally containing it. We note this is a stringent test because the algorithm is not credited for near-misses (as when the correct association is the second most similar concept map). The algorithm performance was compared to a baseline algorithm that constructs its document vector representation by solely based on keyword frequency.

The results of the test showed 27.6% accuracy for the baseline, compared to 43.2% average accuracy for the algorithm. The improvement when a concept map index is used instead of a keyword-based frequency count suggests the value of using the structure of the automatically produced map to identify the most relevant concepts in a document. The results may also benefit from the concept extraction step removing keywords corresponding to linking phrases that would otherwise become noise when comparing with concept labels on the knowledge model test set.

We note that the results produced by the current algorithm do not match the raw performance of state-of-the-art document classification methods, but believe that this difference can largely be accounted for by differences between tasks. Unlike traditional document categorization tasks, for which classes are defined by sets of documents, for this task classes are single concept maps containing limited numbers of concepts, usually no more than 20. Consequently, much less information is available to the system for the classification, and this is only somewhat alleviated by the structural information contained in the concept maps.

6 Related Work

6.1 Integrating external information resources into concept map knowledge models

Recent research has applied novel information retrieval solutions to proactively search the web [20] and specific document libraries [26] for resources that are topically related to a concept map under development. These efforts are complemented by a family of interactive concept suggesters that gather lists of relevant terms to extend a concept map [7, 19]. While these solutions aim to provide online assistance to users for concept map construction, our work focuses on the opposite problem; aiming to use existing expert knowledge models to assist during document understanding and contextualization tasks. The goal of our research is to find useful methods for associating documents to concept maps. Our current approach applies natural language processing techniques to build a concept map representing the document’s information, for use as in index into a family of concept maps.

6.2 Document categorization

Document categorization algorithms [28] have been applied for many years in a wide variety of applications [17]. Most of them use well-known machine learning techniques to find patterns in sets of document features to produce accurate classifiers. Usually, a document is represented by the set of keywords it contains (weighted based on absolute term frequency or on relative frequency in the collection [3]) and an explicit target function is generated, based on a predefined document training set (as when using neural networks or maximum entropy classifiers).

Our approach differs in two ways. First, our document representation is based on concept map fragments as indices, for which we expect structural information to provide a more accurate representation of its content compared to a set of weighted keywords. Structure-based representations of documents have been successfully used in several information retrieval applications [16, 21]. Second, our approach differs in that we are using a lazy evaluation of the classification function, by searching for the most similar element in

the search space (in our case a map in the collection of concept maps), rather than for a fixed target. This approach is more similar to that applied in K-nearest-neighbor and case-based reasoning, but differs from much of that work in that our search space's elements are indexed by structured features.

6.3 Producing concept maps from documents

Some prior work attempts to construct concept maps or similar representations automatically from text. WordNet has been used in to extract a hierarchy of nouns from a document and build a list of concepts, followed by several user feedback iterations to deduce relationships between pairs of concepts and assign initial labels to relations [2]. Another approach relies on a predefined list of domain specific concepts provided by an expert [12]. It considers two concepts to be related if they occur in the same sentence, but does not suggest possible linking phrases. A third alternative focuses on word sense disambiguation [25], using the meaning of nouns and verbs to search for Noun-Verb-Noun structures in the sentences, which become the concept - linking phrase relations.

Our approach is different in that it uses the syntactic structure of the sentences and dependency information to find relations between the words. Also, the relations are not retrieved from predefined ontologies, but are generated from the document itself. This enables the approach to be applied in any domain and makes the results potentially more sensitive to the intentions of the document author. For example, even if two concepts are related in a particular ontology, the author might have intentionally ignored that relation, because it did not correspond to the desired level of abstraction; our approach would not include the relationship in the document description.

In addition, our algorithm produces concepts based on a sentence parser trained to recognize specific syntactic structures rather than on individual words, making the concept labels more complete. Because they are captured from passages in documents, we conjecture that they may also be closer to the concept descriptions produced by a person.

7 Summary and Future Work

We present a novel approach for associating relevant information to an input document by finding the most similar concept map to the topic of the document. Electronic concept maps usually have additional multimedia resources that can also be relevant to a document and can facilitate its understanding. The proposed solution uses existing research for searching a concept map collection and also proposes an algorithm to extract information from an input document to produce a concept map index. Our initial evaluation shows encouraging results. We completed a prototype implementation of the system, which is designed for future interfacing with the CmapTools KEA architecture.

For our future work, we plan to explore two different areas. First, the concept normalization stage will be refined to have a more accurate unification of concepts, not only based on similarity between nouns and adjectives. Additionally, we plan to evaluate the impact of different word normalization strategies, as a way to improve performance. Second, we are planning for a more robust implementation of the system to permit a human subjects study to evaluate the performance of the document-to-concept map transformation process. As part of the ongoing research, we plan to apply the described algorithms on human document understanding tasks, providing dynamic final user interfaces that can take the most advantage of both unstructured documents and structured concept map knowledge models.

References

- [1] ABNEY, S. P. Part-of-Speech Tagging and Partial Parsing. In *Corpus-Based Methods in Language and Speech*, K. Church, S. Young, and G. Bloothoof, Eds. 1996.
- [2] ALVES, A. O., PEREIRA, F. C., AND CARDOSO, A. Automatic Reading and Learning from Text. In *Proceedings of the International Symposium on Artificial Intelligence (ISAI'2001)* (December 2001), pp. 302–310.
- [3] BAEZA-YATES, R., AND RIBEIRO-NETO, B. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

- [4] BASQUE, J., AND LAVOIE, M.-C. Collaborative Concept Mapping in Education: Major Research Trends. In *Proceedings of the Second International Conference on Concept Mapping (CMC 2006)* (2006), A. J. Cañas and J. D. Novak, Eds.
- [5] BRIGGS, G., SHAMMA, D., CAÑAS, CARFF, R., SCARGLE, J., AND NOVAK, J. D. Concept maps applied to Mars exploration public outreach. In *Concept Maps: Theory, Methodology, Technology. Proceedings of the First International Conference on Concept Mapping* (Pamplona, Spain, 2004), A. J. Cañas, J. D. Novak, and F. González, Eds., Universidad Pública de Navarra, pp. 125–133.
- [6] BUYUKKOKTEN, O., GARCIA-MOLINA, H., AND PAEPCKE, A. Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. In *Proceedings of the 10th International Conference on World Wide Web* (2001), ACM Press, pp. 652–662.
- [7] CAÑAS, A. J., CARVALHO, M., ARGUEDAS, M., LEAKE, D., MAGUITMAN, A., AND REICHHERZER, T. Mining The Web to Suggest Concepts During Concept Map Construction. In *Proceedings of the First International Conference on Concept Mapping* (2004), A. J. Cañas, J. D. Novak, and F. M. González, Eds.
- [8] CAÑAS, A. J., HILL, G., BUNCH, L., CARFF, R., ESKRIDGE, T., AND PÉREZ, C. KEA: A Knowledge Exchange Architecture Based on Web Services, Concept Maps and CmapTools. In *Proceedings of the Second International Conference on Concept Mapping (CMC 2006)* (2006), A. J. Cañas and J. D. Novak, Eds., vol. 1, pp. 304–310.
- [9] CAÑAS, A. J., HILL, G., CARFF, R., SURI, N., LOTT, J., GOMEZ, G., ESKRIDGE, T. C., ARROYO, M., AND CARVAJAL, R. CMapTools: A Knowledge Modeling and Sharing Environment. In *Proceedings of the First International Conference on Concept Mapping* (2004), A. J. Cañas, J. D. Novak, and F. M. González, Eds.
- [10] CAÑAS, A. J., LEAKE, D. B., AND MAGUITMAN, A. G. Combining Concept Mapping with CBR: Towards Experience-Based Support for Knowledge Modeling. In *Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference* (2001), AAAI Press, pp. 286–290.
- [11] CHARNIAK, E., AND JOHNSON, M. Coarse-to-Fine n-Best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (June 2005).
- [12] CLARIANA, R. B., AND KOUL, R. A Computer-Based Approach for Translating Text into Concept Map-like Representations. In *Proceedings of the First International Conference on Concept Mapping* (2004), A. J. Cañas, J. D. Novak, and F. M. González, Eds.
- [13] FELLBAUM, C., Ed. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- [14] HARABAGIU, S., MOLDOVAN, D., CLARK, C., BOWDEN, M., HICKL, A., AND WANG, P. Employing Two Question Answering Systems in TREC-2005. In *Proceedings of the 14th Text Retrieval Conference (TREC 2005)* (2005).
- [15] HOFFMAN, R. R., COFFEY, J. W., FORD, K. M., AND CARNOT, M. J. STORM-LK: A Human-Centered Knowledge Model For Weather Forecasting. In *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society* (2001).
- [16] HULTH, A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (2003), pp. 216–223.
- [17] JACKSON, P., AND MOULINIER, I. *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*, vol. 5 of *Natural Language Processing*. John Benjamins Publishing Co, 2002.
- [18] LEAKE, D., MAGUITMAN, A., AND REICHHERZER, T. Understanding Knowledge Models: Modeling Assessment of Concept Importance in Concept Maps. In *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (2004).

- [19] LEAKE, D., MAGUITMAN, A., REICHERZER, T., CAÑAS, A., CARVALHO, M., ARGUEDAS, M., BRENES, S., AND ESKRIDGE, T. Aiding Knowledge Capture by Searching for Extensions of Knowledge Models. In *Proceedings of the Second International Conference on Knowledge Capture (K-Cap 2003)* (2003), pp. 44–53.
- [20] LEAKE, D., MAGUITMAN, A., REICHERZER, T., CAÑAS, A. J., CARVALHO, M., ARGUEDAS, M., AND ESKRIDGE, T. C. “Googling” from a Concept Map: Towards Automatic Concept-Map-based Query Formation. In *Proceedings of the First International Conference on Concept Mapping* (2004), A. J. Cañas, J. D. Novak, and F. M. González, Eds.
- [21] MIHALCEA, R., AND TARAU, P. TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (2004), pp. 404–411.
- [22] NOVAK, J. D., AND GOWIN, D. B. *Learning How to Learn*. Cambridge University Press, New York, 1984.
- [23] NUUTILA, E., AND TÖRMÄ, S. Text Graphs: Accurate Concept Mapping with Well-Defined Meaning. In *Proceedings of the First International Conference on Concept Mapping* (2004), A. J. Cañas, J. D. Novak, and F. M. González, Eds.
- [24] PAPANIKOLAOU, K., GOULI, E., AND GRIGORIADOU, M. Accomodating Individual Differences in Group Formation for Collaborative Concept Mapping. In *Proceedings of the Second International Conference on Concept Mapping (CMC 2006)* (2006), A. J. Cañas and J. D. Novak, Eds.
- [25] RAJARAMAN, K., AND TAN, A.-H. Knowledge Discovery from Texts: A Concept Frame Graph Approach. In *Proceedings of the 11th International Conference on Information and Knowledge Management* (2002), pp. 669–671.
- [26] REICHERZER, T., AND LEAKE, D. Towards Automatic Support for Augmenting Concept Maps with Documents. In *Proceedings of the Second International Conference on Concept Mapping (CMC 2006)* (2006), A. J. Cañas and J. D. Novak, Eds.
- [27] REICHERZER, T., AND LEAKE, D. Understanding the Role of Structure in Concept Maps. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (2006), pp. 2004–2009.
- [28] SEBASTIANI, F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34, 1 (2002), 1–47.
- [29] VALERIO, A., AND LEAKE, D. Jump-Starting Concept Map Construction with Knowledge Extracted From Documents. In *Proceedings of the Second International Conference on Concept Mapping (CMC 2006)* (2006), A. J. Cañas and J. D. Novak, Eds., vol. 1, pp. 296–303.