Concept Maps: Theory, Methodology, Technology
Proc. of the Second Int. Conference on Concept Mapping
San José, Costa Rica, 2006

TOWARDS AUTOMATIC SUPPORT FOR AUGMENTING CONCEPT MAPS WITH DOCUMENTS

Thomas Reichherzer and David Leake, Indiana University, USA
Email: {treichhe, leake}@cs.indiana.edu

Abstract. Concept mapping has been used extensively in educational settings as a learning and teaching tool and, more recently, as a knowledge elicitation tool to capture expert knowledge for preservation purposes. In these contexts, electronic concept maps annotated with supplementary multi-media resources can provide a rich source of information. Unfortunately, it can be challenging for users to identify the right resources to attach, especially when dealing with large resource collections. This paper presents ongoing research on methods for easing the annotation task by automatically searching a document library for relevant documents and suggesting them as potential associations for concepts in a map. The paper begins by discussing how concept map structure can be exploited to automatically generate queries to a database of indexed documents, to search for documents to link to the concepts in a map. It then presents methods for indexing documents to improve the search results. The methods for generating queries and indexing documents have been evaluated using two pre-existing expert knowledge models, with encouraging results.

1 Introduction

Concept mapping (Novak & Gowin, 1984) has been widely used in classroom settings to enable students at many different levels to externalize their knowledge for examination and to aid them in constructing new knowledge by linking new, observed concepts to those already known. Likewise, the naturalness of the concept mapping process can enable domain experts to enter their knowledge directly without the need of a knowledge engineer, and the conciseness and structure of concept maps aids others in understanding the entered information. To facilitate electronic concept map construction and sharing, the Institute for Human and Machine Cognition (IHMC) has developed CmapTools, a suite of publicly-available software tools to support generation and modification of concept maps in an electronic form (Cañas et al., 2004). The CmapTools software enables interconnecting maps and annotating them with material such as documents, images, diagrams, and video and audio clips, providing rich, browsable *knowledge models* available for navigation and collaboration across geographically-distant sites. CmapTools has been used in public outreach programs and large institutional memory and expert knowledge preservation tasks for domains including Mars exploration (Briggs et al., 2004), launch-vehicle systems integration (Coffey, 1999), mesoscale weather forecasting (Hoffman et al., 2001), and nuclear power air effluent analysis (Coffey et al., 2004).

To aid the construction of knowledge models, in collaboration with IHMC, we have conducted a long-term effort to develop support tools that can simplify an expert's task in building rich knowledge models. Our past research efforts have focused on the development of "intelligent suggesters" for content-based support, to aid experts in (1) extending their concept maps with new concepts and propositions and (2) determining topics of new concept maps to be added to the knowledge model (Cañas et al., 2002, Leake et al., 2003, Maguitman et al., 2005). Our current focus is to develop support tools to aid in annotating concepts in concept maps with relevant documents. To build knowledge models, experts must construct concept maps, identify resources from available repositories that they want to link to the model, and decide the specific concepts to which the links should be attached. This task can be daunting if the document libraries are large and if the experts have incomplete knowledge of the information contained in each document. Consequently, automatic methods are needed to select candidate documents from a source document library and to suggest target concepts within the model to connect them via navigational links.

This paper begins with a synopsis of specific motivations for this project, in terms of the potential applications for tools to automatically annotate concept maps with documents. It then briefly summarizes our previous work on cognitive models of concept importance judgments, a foundation on which our concept map annotation work builds, before presenting new methods we have developed for the annotation task. It continues by addressing two research questions: (1) how to exploit the structure of concept maps when automatically generating concept-map-based queries, and (2) how to index text documents to facilitate search for relevant documents to annotate concept maps. As steps towards addressing these

questions, we present results from a Web experiment on the indexing power of concepts and linking phrases, present an algorithm for automatic query generation, and present methods for indexing documents and selecting candidate documents from a document library. We conclude with an evaluation of our indexing and search algorithms using two large-scale expert knowledge models.

2 Motivations for Automatic Document Indexing Methods Using Concept Maps

Tools suggesting documents to link to a concept map's concepts can play a valuable role in aiding users building concept maps, by helping them to supplement knowledge in concept maps with links to relevant documents. This paper introduces the CmapAIDE (Concept Map Automatic Indexing for Document Examination) system, a testbed system integrated into CmapTools to support domain experts in finding relevant documents to link to concepts in concept maps. Currently, the system unobtrusively suggests candidate documents for every concept in a concept map for which it finds relevant documents. We envision extending this process to allow users to specify specific sets of concept of interest, for CmapAIDE to perform focused search.

From a broader perspective, we see tools such as CmapAIDE as a step towards the long-term goal of providing general access to information in document libraries. With electronic concept mapping software such as CmapTools facilitating construction and sharing of knowledge models, we expect that expert knowledge models will become more readily available for people to use, either "as-is" or slightly customize for their personal needs. Such public knowledge models could be used to access and navigate electronic documents in libraries within the context of the concept maps, provided automatic methods exist to make it feasible to annotate large collections of concept maps with documents on demand. This approach to information access will enable individuals to use the captured knowledge in the concept maps to find related information and to help understand the information in the documents within the context of the concept map.

Figure 1 shows the possible use cases. In the first, a domain expert constructs a knowledge model and utilizes automatic indexing methods to annotate the model with supplemental material. In the second, a user downloads an expert knowledge model, customizes the model if needed, by revising, removing or adding concept maps, and then uses the model to access and navigate documents.

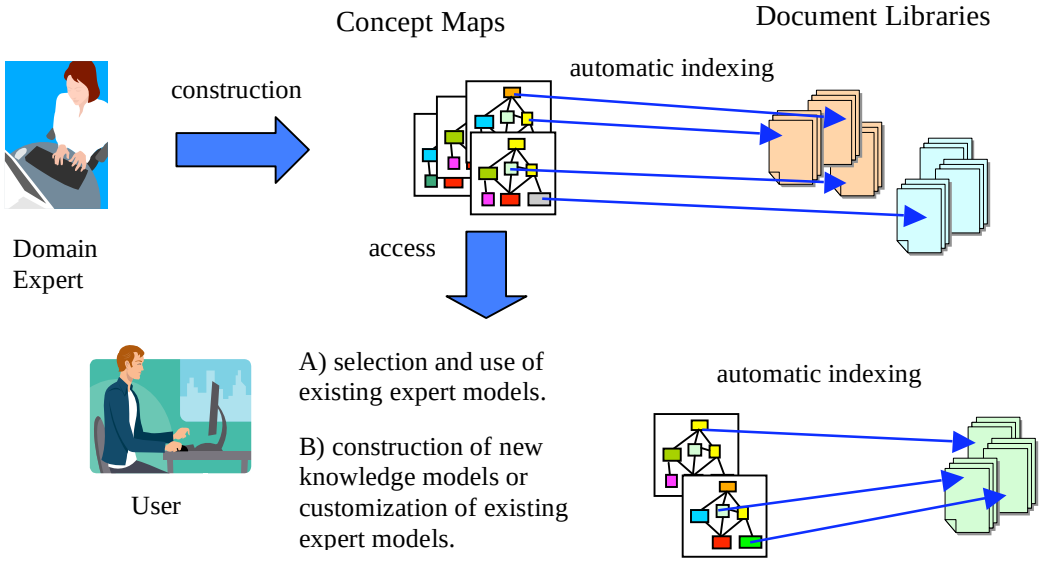

Figure 1: Use cases for automatic indexing.

3 Foundations for the Methods

Our approach to automatic document indexing builds on our previous research on analyzing the relationship of concept map structure and concept importance. Unlike text documents, concept maps have a rich structure that may be exploited by search and navigation tools. In two previous human-subjects studies, we examined how observed structure influences people's understanding of concept importance in already constructed maps and how concept map structure reflects the map-builders' own judgments of concept importance. In the first study, subjects observed concept maps with variations in their topology and layout, including changes in a concept map's number of outgoing and incoming connections, distance to the root concept, and layout differences. When labels of the concepts were replaced with artificial words to exclude domain knowledge about concept importance, structure influenced assessment of concept importance, but layout did not. In the second study, subjects constructed a concept map on a topic of their choice before ranking the importance of selected concepts extracted from their map, whose connectivity or distance to the root concept varied, in order to study how the topology selected by the author related to the author's choice of important concepts in describing the map's topic. Analysis of the results from both studies showed that topology alone is a good indicator (human judgments were closely related to topological factors, with few exceptions), making structure useful to predict people's assessment of concept importance in concept maps and hence to extract topic-relevant information from concept maps. In addition, the studies evaluated the ability of three candidate models to predict concept importance in concept maps based only on structural factors. The best fit was achieved by the Hub-Authority-Root-Distance (HARD) model, which was shown to be sufficient to account for most subjects' importance assignments. The HARD model, shown in Table 1, is used for concept importance weighting in CmapAIDE. The model takes hub (h), authority (a), and upper values of a concept into consideration, as well as three model parameters (α, β, γ) used for tuning, to compute the importance of a concept. Full details on the studies and models are available in (Leake et al., 2004a, Reichherzer & Leake, 2006).

Hub Authority Root Distance (HARD)	$W(C) = (\alpha \cdot a + \beta \cdot h + \gamma \cdot u),\ \gamma \geq 0$

Table 1: Model for assessing concept importance.

4 Retrieval Performance Using Concept Keywords versus Linking Phrase Keywords

The prior research described in Section 3 showed the value of structure in predicting concept importance, and provides a basis for using structural factors to weight concept importance for retrieval. Another important question for generating queries from concept maps is the relative value of keywords extracted from concepts or extracted from links as terms for information retrieval and indexing. Informal results suggest that the prevalence of common linking phrases in concept maps (e.g., *has*, *includes*, *is composed of*) may make linking phrases less informative and less useful for systems that do not employ natural language processing techniques to determine their meaning and the role they play in a concept map. To provide a more definitive result, here we present a new empirical study, expanding on (Leake et al., 2004b), to examine the retrieval power of linking keywords and concept keywords as well as keywords from multiple model-selected concepts. This study involved (1) selecting concepts and links from a concept map, (2) generating Web queries from the keywords of the concept and link labels in different ratios, (3) retrieve Web documents matching the keywords and (4) comparing the documents to the target concept map. To measure similarity between a concept map and a Web document and the retrieval power of keywords from concept maps, we use variants of the standard recall and precision measures, defined with respect to a target concept map *M* and a Web document *D*. Table 2 depicts these measures. The sets *D*, *M*, M_C, and M_L consist of keywords extracted from the Web documents and a concept map respectively, with M_C containing keywords from concept labels and M_L containing keywords from link labels only. *Q* consists of the terms that appear in a query submitted to a search engine, which are considered in the measures to determine the recall and precision values with respect to keywords other than the query terms.

Recall	$R(Q,M,D) = \frac{\lvert(M \cap D) - Q\rvert}{\lvert M - Q\rvert}$	$R_C(Q,M,D) = \frac{\lvert(M_C \cap D) - Q\rvert}{\lvert M_C - Q\rvert}$	$R_L(Q,M,D) = \frac{\lvert(M_L \cap D) - Q\rvert}{\lvert M_L - Q\rvert}$
Precision	$P(Q,M,D) = \frac{\lvert(M \cap D) - Q\rvert}{\lvert D - Q\rvert}$	$P_C(Q,M,D) = \frac{\lvert(M_C \cap D) - Q\rvert}{\lvert D - Q\rvert}$	$P_L(Q,M,D) = \frac{\lvert(M_L \cap D) - Q\rvert}{\lvert D - Q\rvert}$

Table 2: Metrics for measuring precision and recall with respect to a target concept map.

In our study, ten randomly selected concept maps from each of the Mars 2001 (Briggs et al., 2004) and Storm-LK (Hoffman et al., 2001) expert knowledge models were considered to compute Web queries and compare matching Web pages with concept maps. In each experiment, a set of queries was computed, differing in both the ratio of concepts and linking phrases and in the concepts from which the keywords were drawn. The set of combinations considered is shown in Table 3, which reports the results from the study. The query type column indicates the type of query that was generated and submitted to the Google search engine. CCC indicates that keywords were extracted from a concept and two of its connected concepts, LCL indicates that keywords were extracted from a concept and two connected linking phrases, and LLL indicates that keywords were extracted from three linking phrases connected to a concept. For each concept in each map in the knowledge model (provided the concept was connected to more than one other concept) all types of queries were generated. The results were averaged over all concepts considered. All other types of queries considered in the experiment, involve the three topological models. About 25% of the highly ranked concepts as determined by the models were selected to generate Web queries, considering all possible combinations for selecting query keywords. The results were averaged across the different combinations. Recall, calculated as in Table 2, indicates the keywords that were recalled from the Web documents. Precision measures how precisely the document and concept map match. The precision value can be small, especially if retrieved Web documents are large, containing many more keywords than the concept map.

Model	**Query Type**	**R(Q,M,D)**	**R_C(Q,M,D)**	**R_L(Q,M,D)**	**P(Q,M,D)**	**P_C(Q,M,D)**	**P_L(Q,M,D)**
Mars 2001	CCC	0.346	0.345	0.346	0.063	0.049	0.015
	LCL	0.278	0.253	0.372	0.044	0.034	0.011
	LLL	0.192	0.142	0.419	0.025	0.015	0.011
	CCC CRD	0.410	0.414	0.400	0.135	0.110	0.031
	CCC HARD	0.434	0.440	0.417	0.150	0.117	0.034
	CCC PF	0.462	0.471	0.445	0.153	0.121	0.033
Storm-LK	CCC	0.427	0.435	0.418	0.057	0.048	0.010
	LCL	0.317	0.303	0.402	0.053	0.044	0.010
	LLL	0.173	0.142	0.344	0.030	0.022	0.009
	CCC CRD	0.510	0.518	0.507	0.125	0.104	0.024
	CCC HARD	0.454	0.463	0.451	0.114	0.095	0.021
	CCC PF	0.515	0.524	0.507	0.131	0.109	0.025

Table 3: Results of the empirical study to search for related Web documents using concept and linking phrase keywords.

The results support that concept keywords are generally more useful in retrieving relevant Web documents as measured by recall and precision. The recall and precision values for CCC queries are higher than for LCL and LLL queries. It is important to note that queries including terms from linking phrases generally retrieve documents containing terms from other linking phrases, but not as many terms from other concepts. This is particularly notable for queries containing linking phrase keywords only. In addition, searching for relevant Web documents using highly ranked concepts based on our topological models returns better results than using any three directly-linked concepts from a map. The results support the use of concept-based queries, as favored by the models, to generate contextual information for searching related documents. This guides the decision of the design of our automatic query generation method to use concept-based queries.

5 An Algorithm for Automatic Query Generation to Find Relevant Documents

The success of automatic query generation depends on the selection of useful keywords. For the keywords to achieve good recall, the terms in the query must reflect both (1) the target concept to which retrieved documents will be linked, and (2) topic descriptors that support disambiguation of the meaning of the target concept—In concept maps, concepts rely on the context provided by the concepts and connections to define their meaning.

We have developed an algorithm which uses the HARD model as a basis for determining the context keywords of a search query involving a specific concept in a concept map. To generate a query to search for documents relevant to a target concept, first, the top model-favored concepts are selected to build a query context. Keywords from the concepts in the query context are extracted and weighted using a simple frequency model in which each keyword's weight is set to its number of occurrences. The weights of the target concepts are assigned a constant weight. The weighted keywords create a query keyword vector (Baeza-Yates & Ribeiro-Neto, 1999) that can be used to search for relevant documents, indexed with a vector model using a cosine similarity measure. Table 4 summarizes the algorithm used to search for relevant documents.

ALGORITHM:
INPUT:
M: a concept map.
c_t: a target concept
L: a library of indexed documents
τ: threshold for selecting documents from *L*
w_t: the weights of keywords from the target concepts
OUTPUT:
a list *R* of ranked documents *D* from library *L*
BEGIN
//topological analysis
Use the HARD model to assign a weight *W(c)* to each concept *c* in the concept map *C*
Rank concepts according to their weights *W(c)*.
// query formation
Select top 25% of the model-favored concepts; extract keywords to generate a query context.
For each keyword k in query context:
compute keyword weight W(k) as the frequency of k in the query context.
For each keyword k in the target concept c_t:
assign keywords W(k) the fixed weight w_t.
Combine keywords to form a query vector *Q*.
//search
for all documents *D* from *L* do:
compute cosine similarity *S* between *Q* and *D*,
if $S > \tau$, add *D* to list of results *R*.
rank *R* according to similarity *S*, and return it.
END

Table 4: An algorithm for searching relevant documents.

6 Document Indexing to Facilitate Retrieval

Just as concept maps may be analyzed for important concepts, documents may be analyzed to develop indices to facilitate retrievals useful for the task of annotating concept maps. In well-written documents, authors present their thoughts and ideas in a structured form to support readers' understanding. The smallest unit of information tends to be a sentence, followed by a paragraph which discusses a concept, presents an idea, or addresses a question. Large documents are likely to cover a variety of different ideas and topics. Thus, if the goal is to link concept maps to text documents, considering the entire document in the search for relevant information may be less useful than considering smaller units of information.

Automatic text decomposition can aid in finding text *segments* as neighboring units exhibiting internal consistency that can be distinguished from the remaining text. Text relationship maps (Salton et al., 1996) can identify text *themes* as semantically homogenous text units within the document as well as identify relationships between segments and themes and the role they play within the entire document. In addition, to segments and themes, we consider *synopses* as sets of units that relate to many other, non-adjacent units in the document. From themes, segments, and synopses, keywords can be extracted to index documents to support effective access to information in documents. Figure 2 depicts a text relationship map for a sample document consisting of eighteen paragraphs, each considered a separate unit. The solid lines mark neighboring units, forming text segments. Dashed lines indicate themes, consisting of non-adjacent paragraphs that are similar to each other, while dotted lines indicate synopses.

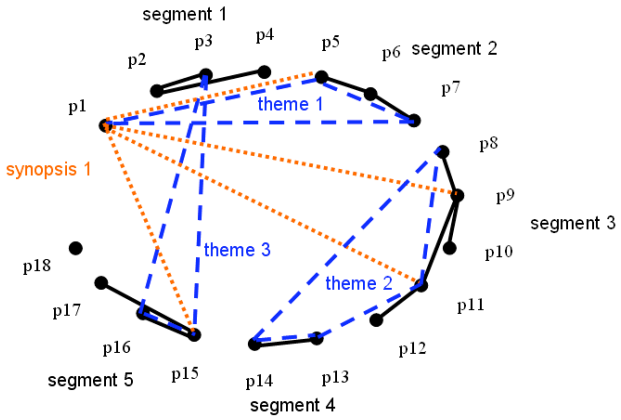

Figure 2: A text relationship map for a sample document.

To compute a text relationship map, we use a keyword vector model computed from the keywords in a unit and apply cosine similarity. The weights for the keyword are computed taking into consideration both the keyword distribution within the units of a document and the keyword distribution throughout the documents of the library, as defined in Table 5. In the formula, the weight of keyword k in document i, considers the frequency of k in i, adjusted by the inverse frequency of k occurring in the units of document i, as well as k occurring in other documents. N_U is the number of units in i and n_{Uk} is the number of times k occurs in units of i. N_L is the number of documents in the document library L, while n_{Lk} denotes the number of times k occurs in documents in L.

$$w_{ik} = \frac{tf_{ik} \cdot \log(N_U / n_{Uk}) \cdot \log(N_L / n_{Lk})}{\sqrt{\sum_j (tf_{ij})^2 \cdot (\log(N_U / n_{Uk}))^2 \cdot (\log(N_L / n_{Lj}))^2}}$$

Table 5: Formula for computing the weight of a keyword in a document.

Once segments, themes, and synopses have been computed, keywords can be extracted and the document indexed by the corresponding keywords and their weights. A document may be indexed multiple times under different sets of keywords corresponding to the segments, themes, and synopses to make it more likely that relevant information embedded in large documents can be identified.

7 Evaluation of the Overall System

We have implemented the methods described in the previous sections in CmapAIDE. The tool pre-processes documents to extract information for comparison with captured knowledge in concept maps. We evaluated the combined indexing and search algorithm using two expert knowledge models that consist of a large number of concept maps linked to several hundred documents. Most of the documents are Web pages, containing text, images, or movie clips. To measure the performance of our methods, we treated the documents annotating the concepts in each expert knowledge model as if removed from the model and collected into a pool of documents for the system to search. For each concept in the original model, which was annotated with one or more documents, we then ran CmapAIDE to test how often CmapAIDE's suggestions included the documents originally chosen. We considered three conditions, in which CmapAIDE generated ranked lists of 5, 10, 20 suggestions. For example, if the expert linked a document entitled "Radiation" to the concept "Human Health & Performance", and CmapAIDE ranks that document eighth on its list of suggestions for that concept, it is a failure for the list of length 5, but a success for the lists of lengths 10 and 20. For the experiment, we varied the influence of contextual information in

retrieving relevant information, considering either both, the target concept (t_c) and topic concepts (T), only topic concepts (T), or only the target concept (t_c).

To compute a document index, using the methods specified above, we focused on applying segmentation to generate an index for the documents; subsequent research will also consider indexing documents based on the extracted themes and synopses. The similarity threshold for computing segments was chosen to be low, favoring large segments in the document. Among the annotation of the knowledge models, only text documents with more than 200 words along with their corresponding concept maps were considered for the performance test. For the Mars 2001 knowledge model (Briggs et al., 2004), performance was tested on a total of 72 concept maps and 159 indexed text documents and for the Storm-LK (Hoffman et al., 2001) knowledge model, the performance test considered 26 concept maps and 95 indexed text documents.

The results show that including the keywords from both topic and target concepts is critical in retrieving relevant text documents with which a concept map may be annotated. When considering lists of 10 to 20 suggestions, CmapAIDE discovered about 70% to 84% of the text annotations as chosen by the experts. When tested with and without indexing documents using segmentation, we recorded an improvement of up to 4% when segmentation is applied. Ideally, we want the system to list the expert's selected text annotations among the top 5 suggestions, facilitating selection. Consequently, subsequent research efforts will focus on improving the selection of the suggestions, considering measures of keyword correlations for closer comparison between the keywords in the target document and concept map. While the current experiment provides a good basis for performance evaluation, we have no complete measure for how many of the suggested annotations may be valid. The experts may have overlooked including some of the text annotations as suggested by CmapAIDE. Thus, the results suggest a lower bound on performance.

Model		**Number of suggestions considered**								
		5			**10**			**20**		
	variations of topic concepts T and target concept t_c in a query	**both**	**no T**	**no t_c**	**both**	**no T**	**no t_c**	**both**	**no T**	**no t_c**
Mars 2001	average matching ratio	0.61	0.25	0.18	0.70	0.31	0.26	0.7	0.35	0.30
	correct suggestions (%)	62.8	30.4	19.4	72.9	35.6	29.6	79.4	38.5	34.4
Storm-LK	average matching ratio	0.57	0.24	0.22	0.75	0.39	0.33	0.87	0.39	0.49
	correct suggestions (%)	48.6	27.5	20.2	69.7	39.4	30.3	84.4	40.4	43.1

Table 6: Results from an evaluation experiment testing performance of CmapAIDE.

8 Summary

This paper presents the design and current results on CmapAIDE, a prototype system to support domain experts in annotating concepts maps with information from document libraries. The same system could also be used in concept-map-based interfaces for navigating documents from a large document library. The system has been evaluated with encouraging initial results. We are continuing to expand and refine the methods to improve system performance and to integrate the system into CmapTools. In addition, we plan to supplement the automated evaluation of the prototype system with a human-subjects study.

9 Acknowledgements

We thank Alberto Cañas and the IHMC CmapTools development team for their continuous collaboration and support.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval, New York: ACM Press.

Briggs, G., Shamma, D., Cañas, A., Carff, R., Scargle, J., and Novak, J. D. (2004). Concept maps applied to Mars exploration public outreach. In Cañas, A. J., Novak, J. D., and González, F., (Eds.), Concept Maps: Theory, Methodology, Technology. *Proceedings of the First International Conference on Concept Mapping*, Pamplona, Spain: Universidad Pública de Navarra.

Cañas, A., Carvalho, M., Arguedas, M. Mining the Web to suggest concepts during concept mapping: Preliminary results. (2002). In *XIII Simpósio Brasileiro de Informática na Educação*, SBIE UNISINOS.

Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Eskridge, T., Gómez, G., Arroyo, M., & Carvajal, R. (2004). CmapTools: A Knowledge Modeling and Sharing Environment. In A. J. Cañas, J. D. Novak & F. M. González (Eds.), Concept Maps: Theory, Methodology, Technology, *Proceedings of the First International Conference on Concept Mapping*. Pamplona, Spain: Universidad Pública de Navarra.

Coffey, J. W. (1999). Institutional memory preservation at NASA Glenn Research Center. Unpublished technical report, NASA Glenn Research Center, Cleveland, OH, USA.

Coffey, J., Eskridge, T., and Sanchez, D. P. (2004). A case study in knowledge elicitation for institutional memory preservation using concept maps. In Cañas, A. J., Novak, J. D., and González, F., (Eds.), Concept Maps: Theory, Methodology, Technology. *Proceedings of the First International Conference on Concept Mapping*, Pamplona, Spain: Universidad Pública de Navarra.

Hoffman, R. R., Coffey, J. W., Ford, K. M., and Carnot, M.-J. (2001). Storm-lk: A human-centered knowledge model for weather forecasting. *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society*, Minneapolis, MN, USA.

Leake, D., Maguitman, A., Reichherzer, T. (2004a). Understanding Knowledge Models: Modeling Assessment of Concept Importance in Concept Maps. *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*, Chicago, Illinois, 795-800.

Leake, D., Maguitman, A., Reichherzer, T. Cañas, A., Carvalho, M., Arguedas, M., Eskridge, T. (2004b). "Googling" from a Concept Map: Towards Automatic Concept-Map-based Query Formation. In Cañas, A. J., Novak, J. D., and González, F., (Eds.), Concept Maps: Theory, Methodology, Technology. *Proceedings of the First International Conference on Concept Mapping*, Pamplona, Spain: Universidad Pública de Navarra.

Leake, D. B., Maguitman, A., Reichherzer, T., Cañas, A.J., Carvalho, M. Arguedas, M., Brenes, S., Eskridge, T. (2003). Aiding Knowledge Capture by Searching for Extensions of Knowledge Models. *Proceedings of K-Cap-03*. Sanibel Island, Florida, ACM Press.

Maguitman, A., Leake, D., Reichherzer, T. (2005). Exploiting Rich Context: An Incremental Approach to Context-Based Web Search. In *Proceedings of Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05)*, Paris, France, Lecture Notes in Computer Science, Vol. 3554, Springer Verlag, pp. 254-267.

Novak, J. D., & Gowin, D. B. (1984). Learning How to Learn. New York: Cambridge University Press.

Reichherzer, T., Leake, D. (2006). Understanding the Role of Structure in Concept Maps. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*, Vancouver, Canada (in press).

Salton, G., Singhal, A., Buckley, C., and Mitra, M. (1996). Automatic text decomposition using text segments and text themes. In *UK Conference on Hypertext*, pages 53-65.