Topic Extraction and Extension to Support Concept Mapping*

David B. Leake, Ana Maguitman, and Thomas Reichherzer
Computer Science Department
Lindley Hall, Indiana University
150 S. Woodlawn Avenue
Bloomington, IN 47405, U.S.A.
{leake,anmaguit,treichhe}@cs.indiana.edu

Abstract

Successful knowledge management may depend not only on knowledge capture, but on knowledge construction—on formulating new and useful knowledge that was not previously available. Electronic concept mapping tools are a promising method for supporting knowledge capture and construction, but users may find it difficult to determine the right knowledge to include. Consequently, knowledge-based methods for suggesting relevant information are desirable for supporting the knowledge modeling process. We are developing methods to aid concept mapping by suggesting relevant information to compare, contrast, and possibly include in knowledge models represented as concept maps. This paper presents two specific methods we are developing for this task, both of which automatically identify topics related to a concept map in order to guide the retrieval of related information. The first, DISCERNER, automatically organizes concept map libraries into a hierarchical structure of topic categories and subcategories that are used as indices for efficient access to relevant stored concept maps. The second, EXTENDER, characterizes the topics of concept maps under construction, applies clustering techniques to the resulting information, and performs incremental web-mining for new but related, topics. It suggests these topics as potential areas for extending the existing concept map or to include in new maps to increase current knowledge coverage.

Key knowledge management issues include how to represent, capture, and construct needed knowledge, and how to access stored knowledge as needed. Concept maps and electronic concept mapping tools provide a representation and knowledge construction method that both novices and experts can use to directly develop and compare knowledge models. Recent research focuses on augmenting these tools by applying knowledge-based technologies to build "intelligent suggesters." These suggesters support users by suggesting relevant concepts, concept maps, and other resources to aid their knowledge construction process.

Intelligent suggesters for concept mapping may draw their suggestions either from existing knowledge models—libraries of previously-defined concept maps—or from external sources. A promising approach to accessing existing concept maps is to apply case-based reasoning methods (Cañas, Leake, & Maguitman 2001; Leake, Maguitman, & Cañas 2002). However, an open issue for these methods is their scalability, which may depend on having a suitable indexing vocabulary to characterize the maps' topics and enable effective retrievals from multiple libraries of concept maps. Because hand-crafting these indices is impractical, automatic index-generation methods are needed. Drawing on external sources, to find related topics that are *not* covered in the system's internal knowledge, also requires identifying topics that are relevant to the areas of current interest, even when they are not available in the current concept map repository. This requires mining external resources, such as the web, to construct knowledge that may aid users in constructing their own knowledge.

This paper describes initial research on two topic extraction methods, one aimed at aiding concept map retrieval, and the other aimed at proposing new topics for a user to consider adding to a concept map or concept map library. DISCERNER uses clustering to build an index enabling efficient search for topic-relevant concept maps, based on the analysis of the map that the user is currently constructing. EXTENDER uses an analysis of the in-progress knowledge model—which may be a concept map under development, or a set of concept maps that capture incomplete knowledge of a domain—to mine the web and construct potentially-relevant topics to suggest to the user.

The paper begins by summarizing the role of electronic concept mapping in knowledge management, the need for intelligent support to aid the concept mapping process, and the usefulness of topic identification in the support process. It then focuses on each of the systems in turn, describing their methods, initial results, and future issues as they are applied to bringing knowledge-based support to the user's knowledge construction and comparison process.

Background

Concept mapping for knowledge modeling: Concept mapping (Novak & Gowin 1984), has been widely used by individuals at many levels—from elementary school students to scientists—to externalize their knowledge, guide knowledge construction, and facilitate knowledge comparison. In concept mapping, subjects construct a two-dimensional, visually-based representation of concepts and

* This research is supported in part by NASA under award No NCC 2-1216. We thank the reviewers for their helpful comments.

their relationships. The Institute for Human and Machine Cognition, at the University of West Florida, has developed CmapTools, a set of publicly-available tools for concept mapping, available at http://cmap.coginst.uwf.edu/. This system is widely used (the tools have been downloaded by users in approximately 150 countries). They support generating and modifying concept maps in electronic form, as well as annotating concept maps with additional material such as images, diagrams, and video clips. They provide the capability to store and access concept maps on multiple servers, to support knowledge sharing across geographically-distant sites. Figure 1 shows a sample concept map of the Mars exploration domain, annotated with links to images and other resources.

The flexible but concise representation of concept maps provides a medium for knowledge capture that is simple enough to be used with minimal training, but that can capture rich conceptualizations in an understandable form, making the concept mapping tools an appealing substrate for knowledge management.

Adding intelligent support: CmapTools provide a framework for knowledge construction, but fully exploiting that framework may be difficult without additional knowledge sources. We are developing an "intelligent suggester" system that explores two techniques for supporting concept mapping. The first applies a knowledge-based method—case-based reasoning—to aid users' knowledge construction and sharing by retrieving relevant concept maps. The second generates new knowledge by mining the web for related topics for the user to consider.

In the case-based retrieval approach, each concept map is treated as a case, with the suggester monitoring the user's concept mapping to suggest concepts included in related concept maps, as well as entire concept maps for comparison. A system has been implemented for this process, with encouraging results for retrieval quality (Leake, Maguitman, & Cañas 2002). However, scaling these methods up to large concept map sets—which in principle could involve all concept map servers using the IHMC tools, around the world—requires new indexing strategies, leading to research on DISCERNER, a system to automatically extract topic information from concept maps.

The success of any case-based approach depends on the existence of previous cases. Consequently, case-based support will have limited usefulness for capturing and constructing knowledge in new domains. This led us to explore methods for bringing an additional knowledge source to bear in aiding the user: Topic knowledge mined from the web. The methods are implemented by EXTENDER, a system which automatically identifies topics to consider including in the current concept map or as a potential topic for a new map.

DISCERNER

When supporting the user by suggesting relevant prior knowledge, rapid retrieval is crucial. DISCERNER (Decision Index for Searching Category Entries by Reducing NEighborhood Radius) uses document clustering to organize concept map libraries into a hierarchical structure of categories and subcategories. Each category contains concept maps involving correlated concepts, with subcategories for smaller, tightly coupled clusters of concept maps within a category. The hierarchical structure serves as an index for efficiently retrieving similar concept maps, using a binary search technique to compare new maps with cluster representatives. Concept maps in leaf nodes are then examined by more expensive methods, resulting in a two-level retrieval strategy.

Building an Index from Concept Maps

DISCERNER uses a greedy agglomerative algorithm to organize concept maps in a hierarchical, tree-like structure. The algorithm starts from a set of initial clusters—clusters containing a reference to one concept map each—and then repeatedly merges clusters whose cluster representatives are closest to each other, by the criterion described below, until all clusters have been merged or the similarities measured between the cluster representatives fall below a pre-set threshold, suggesting that the concept maps from different clusters have little in common and should remain distinct.

During merging process, the algorithm prunes the tree so that only groups of maps above a minimum size form the leafs and inner nodes of the tree, reducing storage requirements significantly and enabling keeping the index in memory, even for large sets of maps. In the result, each tree node marks a category or subcategory with references to concept maps and a cluster representative for later comparison.

A vector-space model for concept map similarity: Concept map similarity is computed from a vector representation of the concept maps. This representation is similar to the popular term-frequency vector with inverse-document frequency adjustment (TF-IDF) (Baeza-Yates & Ribeiro-Neto 1999), but takes advantage of the unique structure of concept maps to adjust term weights. In concept maps, more general concepts are typically found at the top of the map while more inclusive concepts are located at the bottom. The system adjusts weights accordingly, assigning higher weights to keywords from top concepts and lower weights to keywords from concepts located at the bottom of a map. In addition, the system considers the number of outgoing and incoming links to a concept node, strengthening the weightings of keywords in nodes for concepts with many connections to other concepts in the map.

The links in concept maps are considered less important to the categorization of concept maps, and are ignored by the system. More formally, for each concept map c_j of a library of maps $\mathcal{L}$, let $freq_{ijk}$ be the raw frequency of keyword i in concept k of concept map c_j. Assume concept k has n outgoing links and m incoming links and is h steps distant from the top node of the map. The system computes the weight of keyword i of concept k in map c_j as

$$w_{ijk} = freq_{ijk} \cdot (n+m) \cdot \sqrt{1/(h+1)}.$$

The total weight of keyword i in c_j is the sum of all weights w_{ijk} for all concepts k in map c_j. This weight is normalized using the largest keyword weight in concept map c_j and adjusted using the inverse document frequency

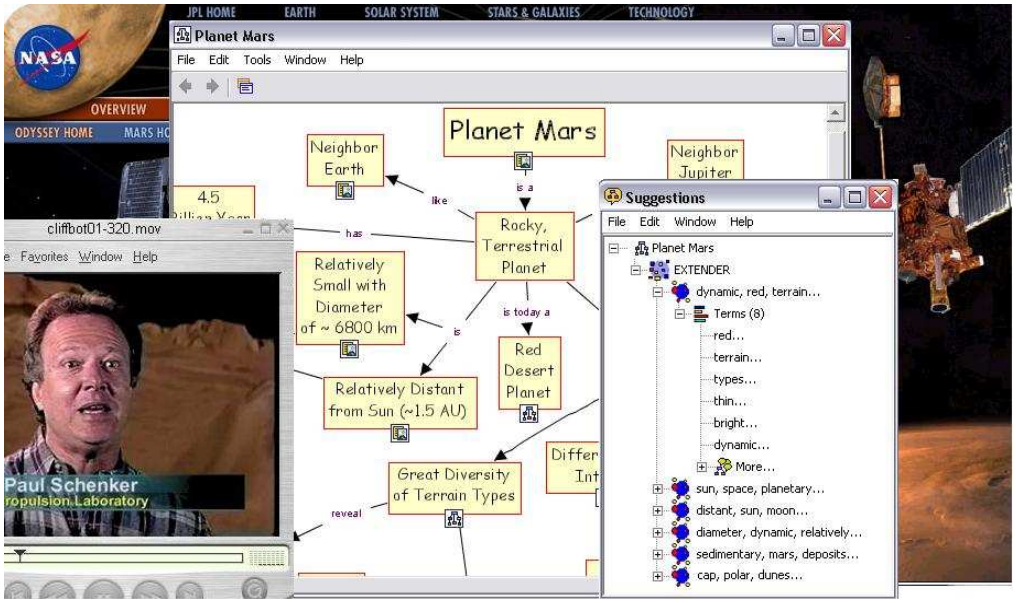

Figure 1: A sample concept map, displayed with the IHMC tools. The EXTENDER suggester window is visible at the right side of the image.

(a concept map is considered to be a single document) for keyword i.

Similarity between these vectors is calculated by the cosine measure. This has a desirable property for centroid-based clustering: When using centroids as the cluster representatives, the inner product of a document with the centroid is the average similarity between the document and all documents in the cluster (Karypis & Han 2000). The clustering algorithm determines the similarity matrix between the cluster centroids, selects the two clusters with most similar representatives as the clusters to merge, and computes the cluster representative of the new cluster as the average weighted sum of the two most similar centroids from the similarity matrix. The end result of the clustering process is a hierarchical categorization indexing all concept maps in the concept map library. We envision applying the index in a central server maintaining a global index of distributed concept map libraries. The server will combine the indices computed independently from the different libraries, using a final clustering step to produce a single hierarchy of categories and subcategories.

Initial evaluation of indexing performance

The concept map index generator has been implemented, and we are evaluating the quality of retrievals based on the index. Our initial experiments tested performance for two data sets. The first contained two knowledge models on similar topics, respectively comprising 93 concept maps from the Mars 2001 project and 9 concept maps on the NASA Centaur Rocket System. The second data set also contained two knowledge models, but on dissimilar topics, with 14 maps on AI topics and 17 concept maps on water and glaciers. The experiments were primarily designed to answer two important questions: (1) After indexing, will DISCERNER find the concept maps it placed into a hierarchy of categories?, and (2) will it merge knowledge models of similar topics into a single category hierarchy while keeping knowledge models from different topics separate? Table 1 summarizes the results from the different tests, for different parameter settings of the algorithm that determine the size of the leaf nodes in the category hierarchy.

	Tests	#Leaf nodes	Error Rate
First Data Set	1	25	17.65%
	2	6	6.86%
Second Data Set	1	9	6.45%
	2	5	3.22%

Table 1: Results from an automatic categorization

For the first data set, the algorithm produces a single category hierarchy that combines the two similar knowledge models, as desired. For the second data set, it produces three different hierarchies in test 1 and two in test 2. The error rate specifies the percentage of concept maps that were unrelated to the category determined by DISCERNER's index. The results show that DISCERNER merges knowledge models from similar topics into a single category hierarchy while keeping knowledge models from different topics separate. In addition, it partitions the data sets into small groups of maps that can easily be analyzed for useful information.

EXTENDER

Research on EXTENDER (EXtensive Topic Extender from New Data Exploring Relationships) combines aspects of *knowledge acquisition*—which assumes that requisite knowledge exists and simply needs to be acquired—

with *knowledge construction*, for a *knowledge extension* approach to knowledge management. In this view, a knowledge model evolves from coordinated processes of knowledge construction and knowledge acquisition. EXTENDER extracts topic descriptions from concept maps, and uses them as the starting point to search for novel related textual information in an iterative process, generating topics at increasing distances from the original map. At each step, the new information is processed, combined with the old, and reorganized to generate descriptions of new related "artificial topics," mined from the web, to suggest to the user.

EXTENDER's topic descriptions are term-based. Initial topics are generated from concept maps (or from a set of concept maps) by extracting keywords from the concept labels in the concept maps, and generating a term vector with weights based on the arrangement of the concepts in the maps. This initial description of the concept map topic is incrementally refined as user and EXTENDER interact. In each cycle, the system takes as input a set of weighted keywords and mines the web—by submitting queries to a search engine (Google), in order to extrapolate a topic from the initial model. In each cycle, a new generation of artificial topics is formed by clustering the results of the web mining process. As this process is iterated, the system focus transfers to novel but still connected topics, with the shift controlled by diversity/focus factors described below. After a number of iterations, the process yields a final generation of artificial topics, which are presented as suggestions to the user. The user can adjust the number of iterations as desired. In general, three iterations appears sufficient to generate a rich variety of artificial topics. The following paragraphs provide details on EXTENDER's approach.

Weighting concept map keywords using topological analysis: Following (Cañas, Leake, & Maguitman 2001), EXTENDER's term weightings summarize the topology of the initial concept maps according to four topological roles, based on algorithms adapted from research on determining hub and authorities nodes in a hyperlinked environment (Kleinberg 1999). *Authorities* are concepts to which other concepts converge. These have the largest number of incoming links from "hub nodes" (defined below). *Hubs* (centers of activity) are concepts with the largest number of outgoing links ending at *authority nodes*. *Upper Nodes* are concepts that appear towards the top of the map in its graphical representation. *Lower Nodes* are concepts that appear towards the bottom of the concept map in its graphical representation. We define four weights, *a-weight*, *h-weight*, *u-weight* and *l-weight*, in [0,1], representing the degree to which a concept belongs to the above categories in a particular concept map.

Combining keywords into queries using concept cohesion: Given a set of weighted keywords describing the topic of a concept map, EXTENDER uses the highest weighted ones to form a web query. Initial queries are constructed exclusively from concept maps using keywords with the highest combined *a-weight*, *h-weight* and *u-weight*; subsequent queries also incorporate new keywords found in web pages. To generate coherent queries, keywords that are adjacent in the analyzed concept maps or documents after stop-word filtering are treated as a unit. Keywords that tend to be part of the same concept or phrase are integrated into a single query.

Term clustering techniques: Textual documents returned by the web queries are analyzed to collect related terms into descriptions of "artificial topics," using a singular value decomposition (SVD) approach. Well-known approaches to information retrieval like latent semantic analysis (Deerwester *et al.* 1990) use SVD to find major associative patterns in the data. This allows to construct semantic indexing mechanisms, where retrieval is not guided by term matching but by nearness at the conceptual space. In order to apply SVD a term-web page matrix is formed from the complete collection of search results. SVD transforms the term-web page vector space into a term-topic and a topic-web page vector space. In the SVD approach, each singular value represents an artificial topic. Principal topics are retained after projecting into a lower dimensional space.

Diversity/Focus factors: The interleaved sequence of search and clustering processes gives rise to generations of artificial topics. These are produced by combining existing keywords with novel keywords from new web searches. In early generations, EXTENDER attempts to diversify its set of terms; in later generations it tries to focus on the new topics found. This is in the spirit of searching and learning techniques (e.g., simulated annealing and reinforcement learning) in which a temperature or curiosity factor is used to favor exploration at the beginning and exploitation during the final stages. EXTENDER uses "diversity/focus factors" to modify the weight of novel and existing keywords. For a novel keyword to be part of a new artificial topic it has to survive a selection process that imposes a minimum weight. The surviving threshold automatically adjusts as a function of the diversity/focus factors.

The EXTENDER Algorithm

The EXTENDER algorithm begins by taking a concept map or a small collection of concept maps about a certain domain, and generates artificial topics as follows:

Step 1: Apply topological analysis to concept maps and use resulting weighted terms to produce the first generation of artificial topics.
Step 2: Use artificial topics aided by concept cohesion to generate queries and present them to a web search engine.
Step 3: Use the diversity/focus factors to integrate the returned results and the old information and complete the term-web page matrix.
Step 4: Apply a term clustering technique to the term-web page matrix and obtain a new generation of artificial topics.
Step 5: Repeat steps 2 to 4 until the final generation of artificial topics is produced.

We are investigating methods for users to introduce bias to constrain this search (e.g., by specifying which artificial topics to consider for subsequent extensions, or by deleting or inserting terms).

Usefulness of EXTENDER's Suggestions

To increase the likelihood that the proposed topics are useful to the user task, it is desirable for the topics to be *coherent*—

Topic 1	Topic 2	Topic 3	Topic 4
mars	climate	robotic	history
science	global	missions	climate
goals	change	human	geologic
exploring	changes	space	earth
nasa	environment	web	change
space	water	exploration	geology
exploration	research	future	processes
missions	activities	system	global
mission	usgs	services	water
system	future	explore	changes
rover	national	nasa	life
future	amp	home	geological
earth	program	page	natural

Table 2: Artificial topics generated starting from a concept map on Planet Mars.

to have a systematic or logical connection among the components that specify a topic; to be *relevant* to the topic at hand, and to build *diversity* by presenting a rich set of topics. As an initial test, EXTENDER was applied to concept maps from the Mars 2001 Project, a library of concept maps created by experts from NASA. For example, from one initial map EXTENDER produced 27 artificial topics, some of which are shown in table 2. Although we cannot present an absolute assessment of the usefulness of the suggested topics, we estimated the system's performance by using the cosine similarity metric for term vectors to compute 2700 similarity values between the 27 suggested artificial topics generated by EXTENDER, and the 100 actual concept maps from Mars 2001 Project, which were not available to EXTENDER but reflect an expert's model of the domain. The similarities from comparing the originating concept map and each artificial topic were always below 0.40, suggesting that EXTENDER is producing novel topics. In 47 instances, similarity measures between new topics and existing maps was greater than 0.50, with 7 greater than 0.70. Thus EXTENDER generated novel topics which often had content similar to that of concept maps created by experts. However, we observed that the artificial topics often overlap. We are now investigating methods for identifying these overlaps, either to integrate similar topics into one topic or to reorganize groups of overlapping topics into more cohesive units.

Perspective

Clustering methods have been widely used to improve the efficiency of information retrieval systems (Rijsbergen 1974; Jardine & Rijsbergen 1971). In DISCERNER, they are used to group concept maps on the basis of the concepts they represent into a hierarchy of categories and subcategories for later retrieval. This relates to case-based reasoning research on partitioning large case-bases into more cohesive sub-case-bases (Yang & Wu 2001) to improve retrieval efficiency and effectiveness. In contrast to the clustering algorithm proposed by *Yang et al.*, our algorithm computes a hierarchical partition that does not rely on a search for optimal parameters to tune the clustering algorithm. The most specific contribution of DISCERNER is its methods to support clustering for concept-map-based case representations.

EXTENDER's approach is novel in identifying related topics using external knowledge sources on the web. The topic characterizations it generates could be considered as indices to probe the memory of the human user, or as indices for *potential* cases not yet stored in the system's case-base. This work is related to the CBR research area of case discovery (McSherry 2000), but is novel in using non-case resources outside the case-base to identify potential topics for previously-unseen cases to solicit interactively.

Conclusion

DISCERNER and EXTENDER illustrate two approaches to automatic topic identification, applied to supporting humans as they build knowledge models. DISCERNER facilitates the use of CBR for concept map management, by organizing concept map libraries into a hierarchical structure of topic categories to aid efficient access. EXTENDER characterizes the topics of concept maps under construction, and mines the web for new but related topics to suggest to the user. Initial results are promising, and we are continuing to evaluate and refine the approaches of both systems.

References

Baeza-Yates, R., and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Reading, MA: Addison–Wesley.

Cañas, A.; Leake, D.; and Maguitman, A. 2001. Combining concept mapping with CBR: Towards experience-based support for knowledge modeling. In *Proceedings of FLAIRS-2001*. Menlo Park: AAAI Press.

Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Furnas, G. W.; and Harshman, R. A. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6):391–407.

Jardine, N., and Rijsbergen, C. V. 1971. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval* 7:217–240.

Karypis, G., and Han, E.-H. 2000. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. TR-00-0016, University of Minnesota.

Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5):604–632.

Leake, D.; Maguitman, A.; and Cañas, A. 2002. Assessing conceptual similarity to support concept mapping. In *Proceedings of FLAIRS-2002*, 168–172. Menlo Park: AAAI Press.

McSherry, D. 2000. Automating case selection in the construction of a case library. *Knowledge-Based Systems* 13(2-3):133–140.

Novak, J., and Gowin, D. 1984. *Learning How to Learn*. New York: Cambridge University Press.

Rijsbergen, C. V. 1974. Further experiments with hierarchic clustering in document retrieval. *Information Storage and Retrieval* 10:1–14.

Yang, Q., and Wu, J. 2001. Enhancing the effectiveness of interactive case-based reasoning with clustering and decision forest. *Computational Intelligence* 14(1):49–64.