Mining the Web to Suggest Concepts during Concept Mapping: Preliminary Results

Alberto J. Cañas, Marco Carvalho, Marco Arguedas

Institute for Human and Machine Cognition

University of West Florida

40 South Alcaniz St., Pensacola, FL 32501, USA

{acanas, mcarvalho, marguedas}@ai.uwf.edu

Summary. The most challenging aspect of constructing a concept map is not coming up with the list of concepts to include, but linking the concepts into meaningful propositions creating a coherent structure that reflects the learner's understanding of a domain. We present an algorithm that, during the process of concept mapping, takes the partially constructed map as input to mine the web, and presents to the user a list of suggested concepts that are relevant to the map under construction. Testing a preliminary implementation of the algorithm with a set of users during a concept-mapping workshop seems to validate its viability. Depending on the size of the suggestion list, the algorithm presented on average between 47% and 69% of the concepts in the final maps before the users added them to the map, showing that the algorithm is able to retrieve concepts relevant to the concept mapping effort.

Keywords: concept map, information retrieval, web mining, meaningful learning

1. Introduction

Concept mapping is a process of meaning-making. It implies taking a list of *concepts* – a concept being a perceived regularity in events or objects, or records of events or objects, designated by a label [NOV 1984], – and organizing it in a graphical representation where pairs of concepts and linking phrases form propositions. Hence, key to the construction of a concept map is the set of concepts on which it is based. In educational settings, teachers often prompt the students by providing an initial set of concepts that they should include in their map.

Coming up with the list of concepts to include in a map is really just an issue of retrieving from long-term memory. In fact, rote learners are particularly good at listing concepts for a domain. It is the process of linking the concepts to create meaningful propositions within the structure of a concept map that is the difficult task. Often, while constructing a concept map, users – whether elementary school students or scientists or other professionals – pause and wonder what additional concepts they should include in their map. It is not that they do not know more about the domain they are modeling, it is that they cannot "remember" what other concepts are relevant.

At the Institute for Human and Machine Cognition (IHMC) of the University of West Florida we have developed CmapTools[1] [CAÑ 2000], a widely-used software program that supports the construction of concept maps, as well as the annotation of the maps with additional material such as images, diagrams, video clips and other such resources. It provides the capability to store and access concept maps on multiple servers to support knowledge sharing across geographically-distant sites.

[1] The CmapTools software package is available for non-for profit use at http://cmap.coginst.uwf.edu.

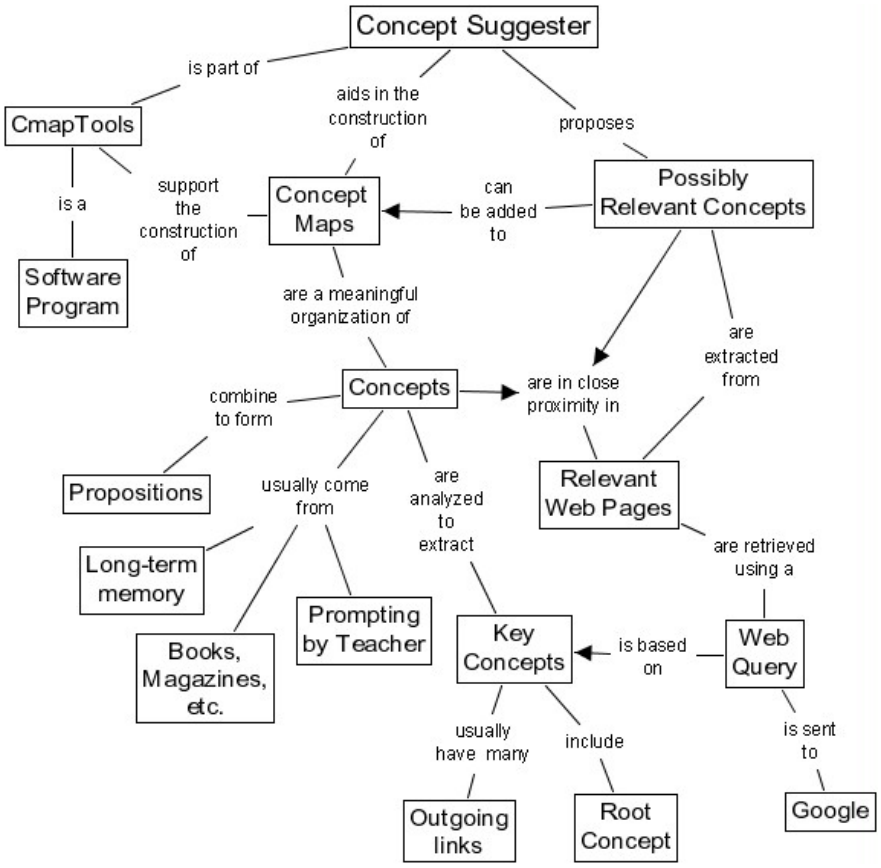

Figure 1: A concept map on the *suggester* module

This paper describes a *concept suggester* module being developed as part of CmapTools that automatically extracts information from a concept map under construction, proactively searches on the World Wide Web (WWW) for concepts that may be relevant to the context of the map, and suggests them to the user for possible inclusion in the map. The relevance of a concept within the process of concept map construction is of course determined by the person creating the map, but we can propose that a relevant concept is one that would likely be added to the map to expand or clarify the knowledge model. This *concept suggester* module is part of a larger effort to aid users in the construction of concept maps. Leake [LEA 2002] describes a module that suggests prior concept maps and associated resources that the user can compare and possibly include as part of the concept map being constructed.

This paper begins with a short description of concept mapping. It then presents the algorithm used to extract relevant concepts from the WWW. Finally, results from an experiment involving professionals during a concept mapping training workshop are presented and discussed. Figure 1 shows a concept map summarizing the purpose and function of the *concept suggester*.

2. Concept Maps and Concept Mapping

Concept maps are tools for organizing, representing and sharing knowledge. Specifically, concept maps, developed by Novak [NOV 1984], have been designed to tap into a person's cognitive structure and externalize concepts and propositions. A concept map is a two-dimensional representation of a set of concepts constructed so that the interrelationships among them are evident (see Figure 1). The vertical axis expresses a hierarchical framework for the concepts. More general, inclusive concepts are found at the highest levels, with progressively more specific, less inclusive concepts arranged below them. These maps emphasize the most general concepts by linking them to supporting ideas with propositions.

Concept maps are assimilation theory's major methodological tool. Ausubel's [AUS 1968] assimilation theory belongs to the family of theories contributing to a constructivist model of human representational processes. Ausubel posits that meaningful learning involves the assimilation of new concepts and propositions into existing cognitive structures. This assimilation of new meaning leads to progressive differentiation and reintegration of cognitive structures. He explicates various forms of meaningful, as opposed to rote learning that involve the assimilation of new information. Ausubel assumes that meaningful learning requires that the learner's cognitive framework contain relevant anchoring ideas to which new material can be related. Indeed, he argues that the most important single factor influencing learning is what the learner already knows. Ascertain this and teach accordingly.

Meaningful learning results when the learner makes a conscious effort to relate new knowledge to be learned with relevant knowledge they already possess. In contrast, rote learning results when the learner memorizes the new information and makes little or no effort to relate and integrate this with their prior knowledge. Information learned by rote is notoriously soon forgotten, and there is little chance for the application of this knowledge in new problem solving contexts [NOV 1998].

There is a growing body of research that indicates that the use of concept maps can facilitate meaningful learning. During concept map construction, meaning making occurs as the learner makes an effort to link the concepts to form propositions. The structure of these propositions into a map is a reflection of his/her understanding of the domain. Therefore, the most important aspect of the meaning making process is not coming up with the list of concepts to include in a map, but establishing the relationship between concepts. A rote learner may very well come up with the same list of concepts as a meaningful learner, but is not able to establish explicitly the relationship between the concepts in the form of propositions. On the other hand, providing a meaningful learner with a richer set of concepts on which to build his/her map can help the learner construct a more complete representation of his understanding of the topic.

3. CmapTools

Software programs like CmapTools make it easier for users to construct and share their knowledge models based on concept maps. In CmapTools we have extended the use of a concept maps to serve as the browsing interface to a domain of knowledge, and provided a tool that allows users to construct, organize, navigate, criticize, and share knowledge models. The software is widely used all over the world, by users who range from elementary school children, to professors creating content for distance learning courses, to NASA scientists. Applications of the tools range from students from different countries collaborating in their knowledge construction [CAÑ 2001] to just-in-time training [CAÑ 1977], to a large multimedia knowledge model about Mars at NASA (e.g. http://cmex.arc.nasa.gov).

4. Suggesting Relevant Concepts

This broad range of users and applications has provided extensive feedback on the process of concept map construction. Taking advantage of this information, we are actively investigating how to enhance the tools with additional features that will proactively aid the users in the construction of their knowledge models. Within this effort, we propose that unobtrusively presenting to the user a list of concepts that seem to be relevant within the context of the concept map being constructed would allow the user to concentrate on the meaning-making process of linking the concepts to form propositions and structuring the map, and away from the effort of "remembering" what concepts are missing.

To find and suggest relevant concepts, we take advantage of various key characteristics of concept maps:

a) Concept maps have structure: By definition, more general concepts are presented at the top with more specific concepts at the bottom. Therefore, different weights can be given to the concepts in the partially constructed maps according to their relative vertical position. Other structural information, e.g. the number of ingoing and outgoing links of a concept, may provide additional information regarding a concept's role in the map.

b) Concept maps are based on propositions: If two concepts form a proposition, the search for relevant documents in the WWW may take into account whether the two concepts appear close together in the text to determine whether the document is relevant.

c) Concept maps have a context: A concept map is a representation of somebody's

understanding of a particular domain of knowledge. As such, all concepts and linking phrases are to be interpreted within that context, and the concept finder can take advantage of it.

5. The *Concept Suggester*

As the user proceeds in the construction of the concept map, the program automatically reviews the changes as they are made and determines when it is appropriate to update the list of suggested concepts. The process of preparing a list of concepts consists of the following steps:

a) Analyzing the partial concept map to prepare a relevant query to use in searching the WWW;

b) Retrieving relevant documents from the WWW;

c) Extracting the relevant concepts from the retrieved WWW pages.

d) Presenting the concepts to the user.

In this section, we describe an initial implementation of steps a) through c) of this algorithm. The purpose of this implementation is to test the viability of the *suggester*. For each of these steps we are aware that significant refinements can be made to improve the relevance of suggested concept, some of which will be discussed at the end of this paper.

This procedure has gone initial testing in a limited environment, and the results are described later in this paper.

5.1. Analyzing the Partial Concept Map

This phase consists of extracting from the concept map a limited set of words that represents its context and that can be used as a query for our meta-search engine.

In traditional information retrieval, word frequency analysis is used to extract keywords from text. This approach, however, would not be effective in a concept map. The concise nature of the map will distort the frequency of words and – furthermore – since in a good map concepts are not repeated, all terms would most likely have the same frequency.

Our approach is to perform a graphical analysis of the partial map to identify the key concepts that play an important role in the context. Specifically, we try to identify concepts that refer to the focus question and concepts that are authority nodes.

Ideally, concepts consist of a single word, or a small set of words. In practice, though, during the process of building a map it is common to find concepts that consist of a large number of words, or even small phrases. For each concept, we try to identify the most relevant words by removing all stop words. If the result is still three words or longer, or if the result is an empty concept, it is discarded for the rest of the process.

At any stage of development, the root node of the map is usually a good representation of the overall topic of the map, or the focus question. Assumed as an important concept, we include the root node as part of our query as long as it consists of less than three words once the stop words are removed.

Authority nodes are those with the highest number of outgoing links to other nodes. We assume that this is an indicative of further elaboration of these concepts, and therefore a gauge of their relevance in the context of the map. The algorithm looks, among all the non-discarded concepts, for those with the largest number of outgoing links. If more than one concept has the same (largest) number of outgoing links, they are all included in the query.

The process then consists of scanning the concept map to locate the root concept and the authority node(s). The overall number of concepts retrieved is dependent on the size of the map. Large maps might have many authority nodes, which would result in a larger number of key concepts, and given the restriction on concepts having less than three words, the process could yield an empty query, in which case the suggester cannot proceed. The query is constructed from the resulting concepts in no particular order.

We plan to enhance this algorithm by performing a noun-phrase analysis of [EVA 1996] of each concept to better identify significant words.

5.2. Retrieving Relevant Documents

We use the query constructed from the key concepts in the previous step to retrieve and rank web pages and build our collection of documents for the concept mining.

We have developed a meta-search engine, based primarily on Google [BRI 1998], in order to retrieve an initial set of documents from the public Internet. The meta-search engine returns a small set of 10 to 20 URLs, depending on the query.

With the documents retrieved, parsed, filtered for stop-words, and indexed, we proceed to the next phase, the actual mining for relevant concepts.

5.3. Extracting Relevant Concepts

Our current approach to extracting relevant concept is simple: search all retrieved documents for all non-discarded concepts from the map. Each time a concept is found in a document, all the neighboring words (excluding stop words) are saved in a temporary table. A word (or noun-phrase) is considered a neighbor if it is part of same sentence as the concept in the text, and is within a specific distance (currently three words) from the concept. In the current implementation, all neighbor words have an equivalent weight and are potential candidates for suggestion. Possible enhancements to the algorithm include filtering this list of neighbors further to extract only noun and noun-phrase candidates.

The result of searching for all of the map's concepts in all the documents is a large collection of terms that are neighbors of the map's concepts in the text. We now proceed to rank these terms using frequency analysis to obtain an ordered list of suggested concepts. The *suggester* can determine the size of the subset of terms to display to the user.

6. Experimental Procedure

During a concept-mapping workshop at IHMC, we used a special version of CmapTools that kept track and logged the users' changes to the map as they built them. This version of the tool did not include the *suggester* module.

We collected logs from seven users during the development of a map with the focus question: "How do we produce electricity?" All users were professionals, experts in the field, and had no previous concept mapping experience.

From each log, we decomposed the concept map construction into a series of steps, each of which usually included the addition or deletion of a concept. Then, for each step, we took the partially constructed concept map and generated, using the algorithm described earlier, the list of relevant concepts that the *suggester* module would have generated at that point had it been part of the software tool. We repeated this process for each step of the construction of each of the seven maps.

Since there was no *suggester* module in the program used, and therefore no list of suggested concepts was presented to the users, there is no way to tell whether they would have taken advantage of the suggested concepts. But we can determine whether, at any step, any of the concepts that the *suggester* listed were added by the user in a subsequent step. For any step in the construction of a map, if the user added concepts that were part of the list prepared by the *suggester* for this step or a previous step, we can conclude that the user considered these added concepts relevant.

7. Experimental Results

Figure 2 shows a partially constructed concept map for subject number 2, together with the list of 15 suggested concepts on the right. (We selected this subject since the data from his/her map was close to the average, as shown in Figure 3.) The top list of 3 concepts (*solar, electricity* and *wind*) includes those that the user added in a subsequent step. The other 12 concepts in the lower list were not included in the map in the subsequent steps. In the concept map itself, the 4 boxes with rounded corners indicate concepts (*heat generation*, *fuel sources*, *nuclear*, and *gas*) that had appeared in the *suggester's* list before they were added in a previous step. As can be seen, close to half of the concepts in the map at this stage had already been suggested. (In fact, since some of the concepts are composed of two words, in many cases they correspond to two suggestions, e.g. *fuel sources*).

The list of concepts suggested but not included in the map deserves some comments. First, there are concepts in the list that do not seem relevant at all (e.g. *Portugal*, *Scotia*, *united*). Other concepts, however, may have allowed the user to improve or expand the map: e.g., *power*, *renewable*, *oil*, *energy*, *environmental, technology*.

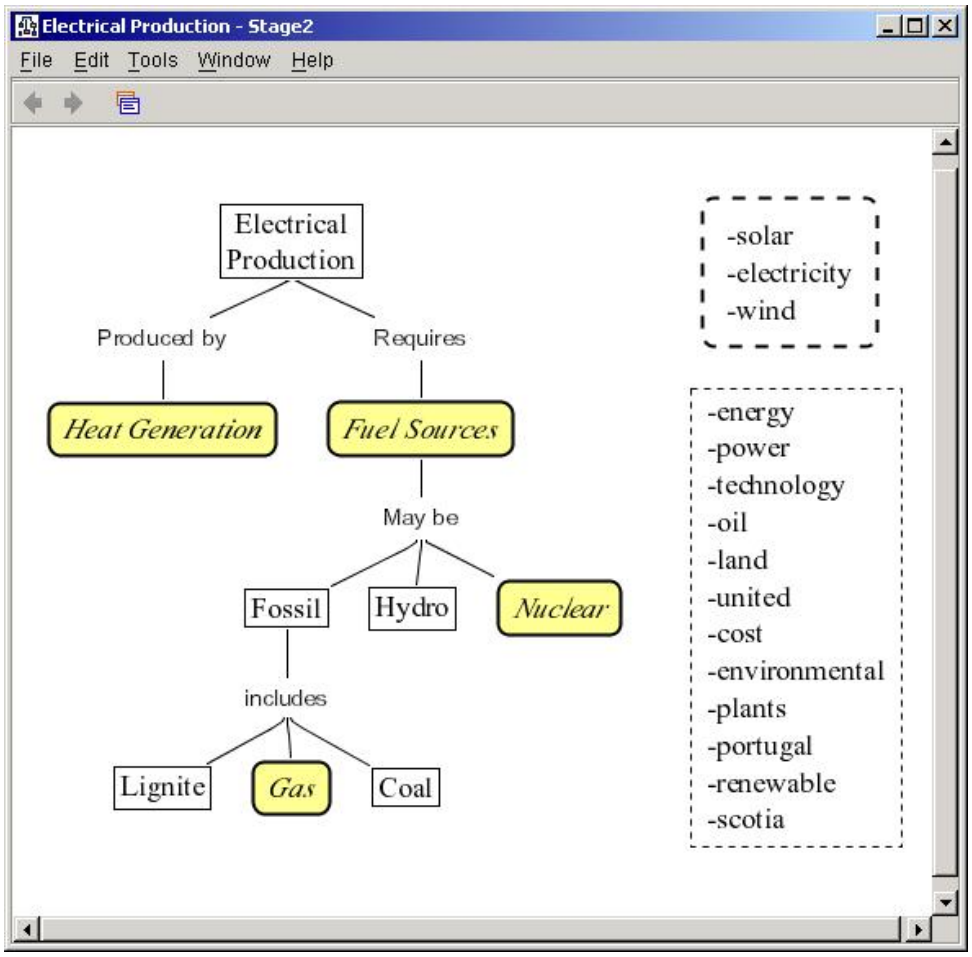

Figure 2: Concept map under construction with suggested concepts included and suggested concepts not included in subsequent steps.

Table 1 shows a summary table for the final concept maps prepared by the 7 subjects. Column 2 shows the total number of unique concepts in each of the 7 maps. (In some maps, the same concept appears more than once – we count these as one concept). In addition to a suggestion list of 15 concepts as shown in Figure 2, the table shows the results for suggestion lists of 25, 50 and an unlimited number (usually between 1000 and 3000) of concepts.

For each of the list sizes, the table presents the number of concepts in the suggestion list that appear in the final map, and that were suggested before the user inserted them in the map. For example, for subject 1, out of 15 concepts in the final map, 9 (60%) were suggested in the list of 15 in a step before they were inserted in the map, and the number goes up to 13 (87%) if we consider the complete list of suggestions.

For the list of size 15, the percentage of concepts already suggested ranges from 23% (subject 4) to 71% (subject 6), with an average of 47%. This means that, on the average, almost half of the concepts in the concept maps would have been displayed by the *suggester* before they were inserted in the concept map by the user. If the list of suggested concepts is increased to 25, the average increases to 56%; if the list is 50 concepts long the average is 65%, and the final column shows that 69% of the concepts used in the map were previously extracted by the program.

We have mentioned previously that in many cases two words from the suggested list formed a

Final Concept Maps

Subject	Total # of Unique Concepts	Number of Concepts in Map that were in a Previous Suggestion List							
		From list of 15 suggestions		From list of 25 suggestions		From list of 50 suggestions		From the complete List	
		#	%	#	%	#	%	#	%
1	15	9	60%	11	73%	12	80%	13	87%
2	16	6	38%	8	50%	10	63%	10	63%
3	16	9	56%	9	56%	12	75%	13	81%
4	13	3	23%	3	23%	3	23%	4	31%
5	28	7	25%	10	36%	13	46%	14	50%
6	17	12	71%	14	82%	15	88%	15	88%
7	13	7	54%	9	69%	10	77%	11	85%
Average	**16.86**	**7.57**	**47%**	**9.14**	**56%**	**10.71**	**65%**	**11.43**	**69%**

Table 1: Number of concepts in the final map that had been suggested in a previous step.

composite concept (e.g. *heat exchange*) in the map, and that we count this as one concept from the suggested list instead of two (which actually reduces by one the size of the list, but we have not taken this into account in the table).

8. Discussion

The results presented suggest that it is feasible to use the WWW to mine for concepts that may be relevant to a user during the process of concept mapping. Even with a small list of 15 suggestions, by the time the concept maps were completed on the average 47% of the concepts in the map would have been suggested before they were inserted in the map. The more concepts in the suggestion list, the higher the percentage. For the complete list of retrieved words, the percentage reaches 69%. This suggests that care must be taken in tailoring the algorithm that ranks the retrieved words to make sure the most relevant are presented.

An argument could be made that there is no use in presenting the learner with concepts that he or she was going to include in the map anyway. However, since there is no reason to believe that the concepts that the user included are the only relevant ones in the list, we can speculate that among the other proposed, the user would have found other concepts that could have enhanced his or her map. In Figure 2 we observed that in the list of concepts suggested, many of those that were not included in subsequent steps would have been useful in extending the map.

The algorithm presented can be enhanced with the expectation of improving performance. Among these changes, we need to perform a better analysis of the concept map to yield a more precise query. At this time, we are not taking advantage of the linking phrases, only the concepts. We only retrieved up to 20 pages from Google; retrieving more pages from several search engines would provide more documents to mine. Indexing the web pages retrieved will allow users to take advantage of web pages retrieved during the map construction process by other users. (An initial test on running the queries for the suggestions several times for all users and accumulating the results yielded an increased final average value of 57% for concepts used from the list of size 15). We are implementing a web crawler that will take the web pages retrieved and search for other pages that are linked to these, improving the cache of pages indexed. Finally, the algorithms used to rank the pages, search for the suggested concepts within the pages, and rank the resulting concepts are also being improved.

To reduce the noise in the concepts suggested, we must ensure that the documents in our collection surpass a minimum level of relevance to the concept map, which is done though our ranking process. Ranking web documents in terms of relevance to a concept map is key to our current research effort. In Carvalho [CAR 2001] we reported successful results by using a matrix comparison analysis between the text and the map. This approach leverages the structure of the map to estimate relevance in text documents from an

analysis of word frequency and proximity, and results in more than a relative ranking among documents. It also provides metrics that allows us to estimate relevance of each page independently. This is an important factor we will use to threshold our collection and preserve only the most relevant documents.

9. Conclusions

The preliminary results presented show that during the construction of a concept map, taking advantage of the structure and semantic of the map we can mine the WWW for concepts that are relevant to the map in progress. Even though in the experiment conducted the users did not rank the concepts being suggested, the fact that the *suggester* was able to present on the average almost half of the concepts before they were used leads us to believe that the list of concepts suggested might include other words that would allow the user to enhance the map under construction. At a minimum, the results justify continuing the development of the *suggester* program in order to test a complete implementation of the module.

10. Acknowledgements

This work was partially funded by NASA's Intelligent Systems Program and by the US Navy's Chief of Naval Education and Training. We would like to thank David Leake, Thomas Reichherzer, Ana Maguitman and Tom Eskridge for their collaboration, the IHMC CmapTools development group for their technical support, and the anonymous referees for their suggestions.

11. References

[AUS 1968] Ausubel, D. P., *Educational Psychology: A Cognitive View*, New York: Holt, Rinehart and Winston, Inc, 1968.

[BRI 1998] Brin, S. and P. Lawrence, *The Anatomy of a Large-Scale Hypertextual Web Search* Web. 7th WWW Conference, 1998.

[CAÑ 1977] Cañas, A. J., J. W. Coffey, T. Reichherzer, N. Suri, R. Carff, and G. Hill, *El-Tech: A Performance Support System with Embedded Training for Electronics Technicians*, Proceedings of the Eleventh Florida Artificial Intelligence Research Symposium, Sanibel Island, Florida, May 1997.

[CAÑ 2000] Cañas, A. J., K. M. Ford, J. W. Coffey, T. Reichherzer, N. Suri, R. Carff, D. Shamma, G. Hill, and M. Breedy, *Herramientas para Construir y Compartir Modelos de Conocimiento Basados en Mapas Conceptuales,* Revista de Informática Educativa, Vol. 13, No. 2, pp. 145-158, 2000.

[CAÑ 2001] Cañas, A. J., K. M. Ford, J. D. Novak, P. Hayes, T. Reichherzer, and N. Suri, *Online Concept Maps: Enhancing Collaborative learning by Using Technology with Concept Maps*. The Science Teacher, 68(2):49-51, April 2001.

[CAR 2001] Carvalho, M., R. Hewett, and A. J. Cañas, *Enhancing Web Searches from Concept Map-based Knowledge Models*, SCI Conference Orlando, 2001.

[EVA 1996] Evans, D. A. and C. Zhai, *Noun-Phrase Analysis in Unrestricted Text for Information Retrieval*, 34th Annual Meeting of ACL (ACL-96), 1996.

[LEA 2002] Leake, D., A. Maguitman and A. J. Cañas, *Assessing Conceptual Similarity to Support Concept Mapping*, Proceedings of the 15th International FLAIRS Conference, Pensacola Beach, FL, May 2002.

[NOV 1984] Novak, J. D. and D. B. Gowin, *Learning how to Learn*. New York: Cambridge University Press, 1984.

[NOV 1998] Novak, J. D*., Learning, Creating , and Using Knowledge: Concept Maps as Facilitative Tools for Schools and Corporations*. Mahwah, N.J.,Lawrence Erlbaum & Assoc,1998.