Concept Mapping: Connecting Educators
Proc. of the Third Int. Conference on Concept Mapping
A. J. Cañas, P. Reiska, M. Åhlberg & J. D. Novak, Eds.
Tallinn, Estonia & Helsinki, Finland 2008

ASSOCIATING DOCUMENTS TO CONCEPT MAPS IN CONTEXT

Alejandro Valerio & David B. Leake, Indiana University, U.S.A
Alberto J. Cañas, Institute for Human and Machine Cognition (IHMC), U.S.A
Email: {valerio, leake}@cs.indiana.edu

Abstract. To be useful, automatic document classification systems must accurately place documents in categories that are meaningful to users. Because concept mapping externalizes humans' conceptualizations of a domain, concept maps provide meaningful categories for organizing documents. Since electronic concept-mapping tools provide mechanisms for using concept maps for effective document access, using concept maps as means to classify documents provides at the same time a browsing system to access the classified documents. To enable automatically associating documents with the relevant concept maps, this paper presents a new top-down/bottom-up approach to classifying documents in the context of topically relevant concept maps. Using the target concept maps as context for extracting concepts from text, this approach generates concept-map-based indexing structures from documents and then indexes them under the concept map most compatible with the document. An experimental evaluation shows marked improvements in performance compared both to a previous bottom-up approach to this classification task and to a second baseline method using unstructured keyword-based indices.

1 Introduction

Automatic document classification is as a powerful tool to help people select and understand relevant documents, by placing documents in the context of topically related information. Electronic concept mapping tools such as the CmapTools suite (Cañas *et al*. 2004), provide an easy-to-use method for humans to generate rich structured descriptions of their conceptualizations –which can in turn be viewed as descriptions of topics of interest– and are widely used for browsing and sharing knowledge. Consequently, the development of tools to automatically associate documents with relevant concept maps would be useful both for helping people to find documents related to a topic of interest as they browse concept maps, and for helping people to understand documents, by suggesting relevant concept maps to provide additional information as they read documents.

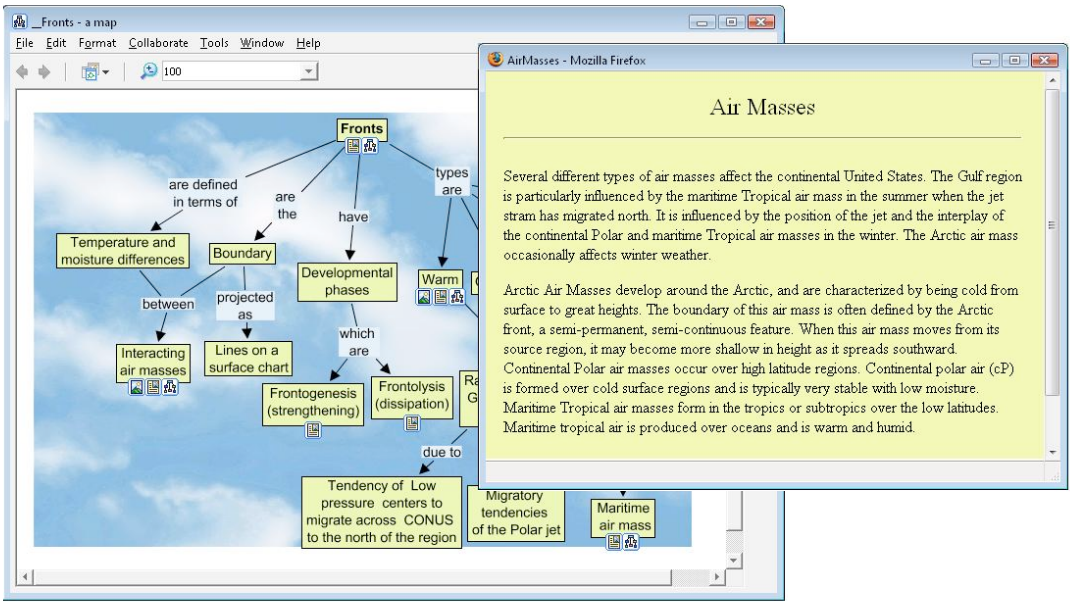
In previous work (Valerio, Leake, & Cañas, 2007), we presented initial steps on a method for document classification in which documents are associated with concept maps, based on comparing the target concept maps to a set of concept map fragments generated automatically from the document, and presented an evaluation demonstrating the promise of that approach. The fragmentary concept maps were generated entirely bottom-up from the text in documents, without considering the set of target concept maps. This paper explores a new top-down/bottom-up approach, which exploits the context of a set of target concept maps to bias assignment of labels for concepts, in an algorithm for extracting concepts from documents. Instead of building a single representation for each document, the approach builds a family of representations; each one optimized for the context of a different target concept map, and then classifies the document by the concept map that generates the best-customized fit. We hypothesize that by using top-down guidance from each map when each index is generated, the resulting sets of concepts map will more closely resemble the concept maps defining the categories, and that this will increase classification accuracy.

The paper begins by describing concept maps and the use of electronic concept maps as a medium for knowledge construction and sharing. It then surveys some related work on associating documents to concept maps, frames our specific problem, and presents our algorithm. Finally it presents an evaluation comparing the new algorithm to the previous algorithm for generating concept-map-based indices, and to an additional baseline using only unstructured keyword-based indices, with encouraging results.

2 Concept map Knowledge Models as a Rich Context for Documents

Concept maps express concepts and relationships in a two-dimensional network, where nodes correspond to concepts and links correspond to concept relationships. Concept mapping was developed in the context of education (Novak & Gowin 1984), but more recently, it has been recognized as a useful tool for knowledge construction and sharing by domain experts. In contrast to formal network knowledge representation models, such as semantic networks, conceptual graphs, and text graphs, concept maps are described in informal terms; they use natural language for concept and link labels, and the concept-link-concept triples form simple natural language propositions.

The CmapTools concept mapping software (Cañas *et al*. 2004) from the Institute for Human and Machine Cognition (IHMC) provides a means for generation and sharing of electronic concept maps, and permits the construction of concept-map-based knowledge models which are collections of topically related linked concept maps with attached resources such as documents or images (e.g., Briggs *et al*. 2004). Figure 1 shows a concept map and a linked document resource as displayed by CmapTools. The rich knowledge provided by the concept map and associated resources is a useful context for human document understanding, if documents can be associated with the proper concept maps. The CmapTools system provides methods for annotating concept maps with documents by hand. However, for document sets that are too large to process by hand, or for automatically monitoring a document stream to suggest documents relevant to topics of interest (already captured in a concept map), it is desirable to develop automatic classification methods.


_Fronts - a map
File Edit Format Collaborate Tools Window Help
100
Fronts
are defined in terms of
are the
have
types are
Temperature and moisture differences
Boundary
Developmental phases
Warm
between
projected as
which are
Interacting air masses
Lines on a surface chart
Frontogenesis (strengthening)
Frontolysis (dissipation)
due to
Tendency of Low pressure centers to migrate across CONUS to the north of the region
Migratory tendencies of the Polar jet
Maritime air mass
AirMasses - Mozilla Firefox

Air Masses

Several different types of air masses affect the continental United States. The Gulf region is particularly influenced by the maritime Tropical air mass in the summer when the jet stram has migrated north. It is influenced by the position of the jet and the interplay of the continental Polar and maritime Tropical air masses in the winter. The Arctic air mass occasionally affects winter weather.

Arctic Air Masses develop around the Arctic, and are characterized by being cold from surface to great heights. The boundary of this air mass is often defined by the Arctic front, a semi-permanent, semi-continuous feature. When this air mass moves from its source region, it may become more shallow in height as it spreads southward. Continental Polar air masses occur over high latitude regions. Continental polar air (cP) is formed over cold surface regions and is typically very stable with low moisture. Maritime Tropical air masses form in the tropics or subtropics over the low latitudes. Maritime tropical air is produced over oceans and is warm and humid.

Figure 1. Example of a concept map and an attached document resource as displayed by CmapTools, from the STORM-LK knowledge model (Hoffman *et al*. 2001) .

An automated procedure to extract information from documents to produce concept-map-based indices must be able to recognize meaningful phrases for concepts and links in input documents in natural language. However, because concept maps are an informal representation, generating a “human-like” concept map, for use as a categorization index to compare to human maps, does not require complete analysis of the meaning of the documents. This makes the associated NLP problem somewhat less complex than full understanding.

3 Prior Work on Associating Documents to Concept Maps

The combined top-down/bottom-up approach contrasts with most prior research on automatic methods to form associations between documents and concept maps, which address the problem exclusively top-down. For example, recent research has applied information retrieval solutions to proactively search the Web (Leake *et al*. 2004) and to search specific document libraries (Reichherzer & Leake 2006a) for resources that are topically related to a concept map under development. However, these solutions aim to provide assistance to users during concept map construction, so the only information that these approaches use from documents is their keywords matching the labels in the target concept map.

Some prior work has instead explored bottom-up approaches, attempting to construct concept maps (or similar representations) automatically from text, but ignoring the information that is available in the possible target concept map knowledge model. Valerio, Leake & Cañas (2007) and Valerio & Leake (2006) apply information extraction techniques to produce a normalized list of concepts, for which labels are assigned by selecting the shortest available label extracted from the document. Alves, Pereira, & Cardoso (2001) use WordNet to extract a hierarchy of nouns from a document and build a list of concepts, followed by iterations of user feedback to identify relationships between pairs of concepts and assign initial labels to relations. Another alternative focuses on word sense disambiguation, using the meaning of nouns and verbs to search for Noun-Verb-Noun structures in the sentences (Rajaraman & Tan 2002). One step towards a more combined approach relies on a predefined list of domain-specific concepts provided by an expert but only considers two concepts to be related if they occur in the same sentence (Clariana & Koul 2004).

4 Overview of the Approach

We address the classification problem starting with a predefined set of concept maps, which constitute the classes. We assume that this set of concept maps will have been generated by hand, by experts or other users, and that the number of concept maps is comparatively small. However, most proposed processing steps are relatively efficient, and some intermediate calculations on the concept map collection can be done offline and stored along with the corresponding map to increase efficiency. In particular, the calculation of the importance of concepts in a map can be executed in this fashion.

The task is to assign each document to the most relevant member of the set of concept maps. Our approach begins by generating sets of indices for each document, each one generated in the context of a different target concept map, in order to bias index generation towards maximizing similarity with the target map. The concept map whose index best matches the corresponding document index is selected as the classification.

More specifically, to associate documents with concept maps, the system takes as input a document and a set *S* of concept maps (called *context concept maps*). For each concept map in this set, the system applies the index generation algorithm (described in a following section) to produce a set of *concept map indices* from the document, in context of that map. This produces n slightly different sets of concept map fragments as the document index. Each document index *index*(*D,C*) makes the concept labels in the index as similar as possible to the labels in the corresponding context concept map *C*, and the concept map most similar to the index is selected. Thus:

$$Class(D,S) = \arg\max_{C\in S} Sim(index(D,C),C)$$

Our approach differs from traditional document categorization algorithms (Sebastiani 2002) in two ways:

1. *Concept map fragments as indices*: Our document representation is based on concept map fragments as indices. The significance of this approach is that these concept map fragments include structural information about concept relationships, which we expect to provide a more accurate representation of its content compared to a set of weighted keywords, and also to enable more effective matching when comparing documents to concept maps, which themselves are structured.
2. *Focus on finding the most similar classification*: Our aim is not to make a boolean decision about whether a document fits a specific fixed category, but rather to identify the most similar element in the search space. This method is in the spirit of K-nearest-neighbor and case-based reasoning, which take a lazy learning approach to categorization. This approach is suitable, for example, when automatically associating documents to the most relevant knowledge model, for a user to make the final determination of whether to add them to the knowledge model.

5 Automatic Generation of a Concept Map Index

Many natural language processing techniques exist for exploiting the information contained on the structure of sentences and phrases of documents (e.g., Harabagiu *et al*. 2005; Alves, Pereira, & Cardoso 2001). For our task of associating documents to existing concept maps, many of the same methods are relevant and could be applied to refine the process. Here we focus on the characteristics of the process which are specific to the task of mapping documents to concept maps.

Our approach revises our previous bottom-up model of concept map generation (Valerio, Leake, & Cañas 2007). That constructed concept maps based solely on the concepts and linking phrases found in the input document. Our central addition is in the *Concept labeling* step, which now assigns concept labels based on the existing labels from an input concept map, to provide a context to bias the map generation. In this way, if a relevant target concept map is known, the labels of the new map may be biased towards the vocabulary used in the target map.

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

Parsing: The document is first preprocessed by a sentence boundary detection algorithm based on regular expressions, followed by a part-of-speech tagger. 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 parsing approach is a modification of Abney's partial parser (Abney 1996) as detailed in (Valerio, Leake, & Cañas 2007).

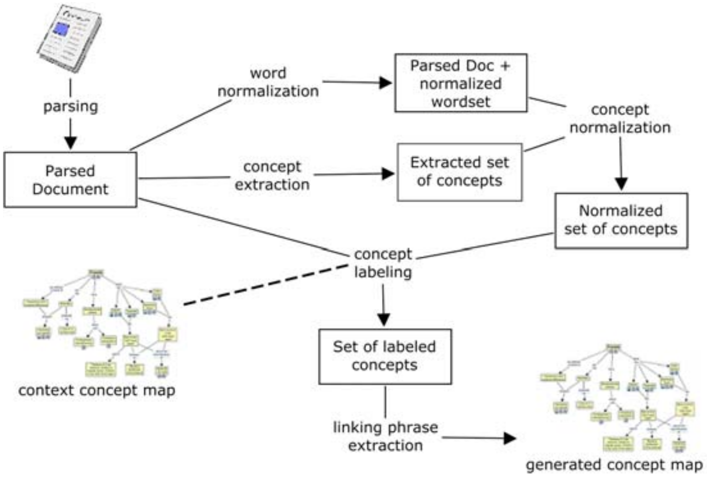


Figure 2. Procedure to construct a concept map index automatically from a document.

Word normalization: Documents contain morphological variations of words that refer to the same entity, and may use multiple synonyms. The word normalization step splits words into disjoint equivalence classes, using a lemmatizer to find the root of the words (e.g., the root word of “realizing” is “realize”), and a part-of-speech tagger and WordNet (Fellbaum 1998) to find synonymy relations. Once the algorithm identifies the word equivalences, it tags each word with its class, for use in comparing words in later steps.

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

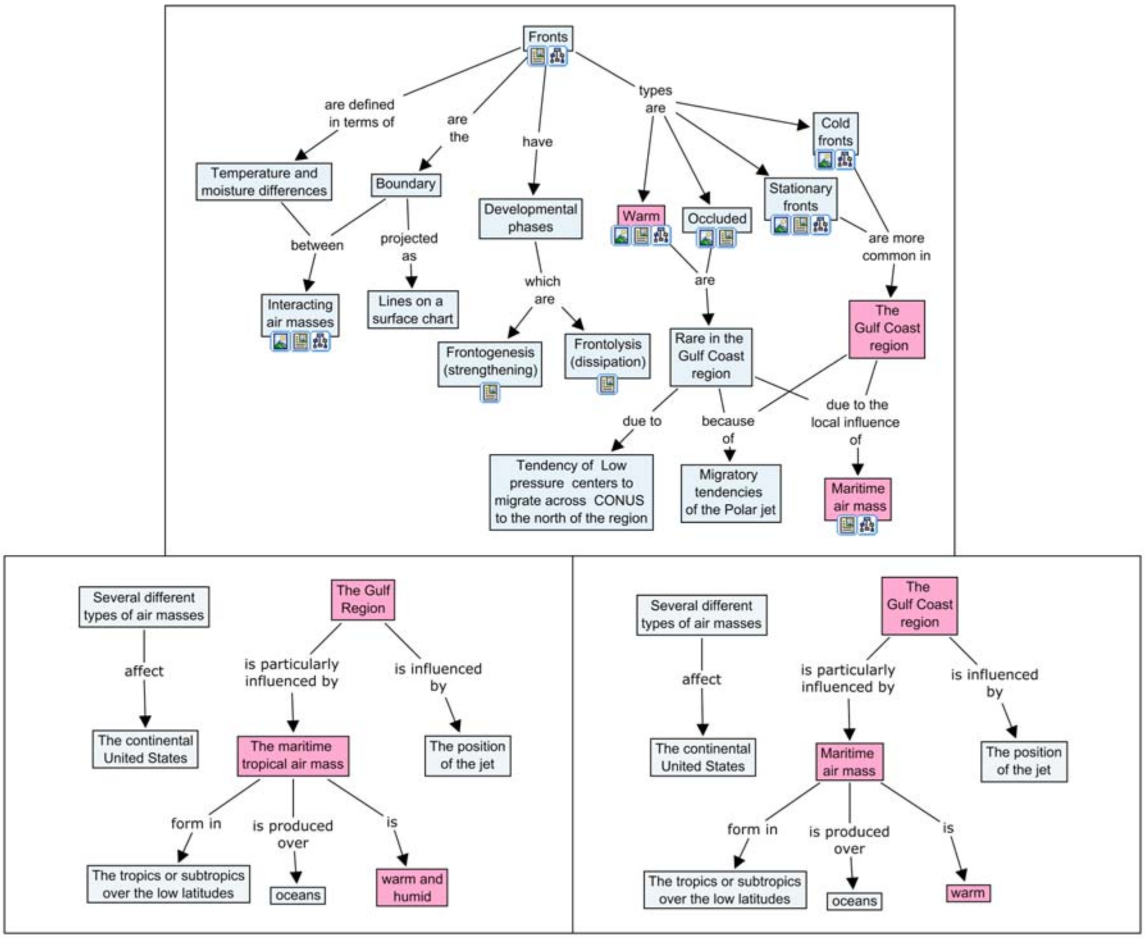
Concept normalization: The sentence chunks corresponding to concept labels may have superficial differences despite some of them referring to the same concept. The normalization step implements a simple solution for co-reference resolution. Two concept labels are considered the same if all nouns and adjectives in either one are contained in the other, considering the classes produced during the word normalization step. This procedure is applied to resolve named entity co-references as well. The primary challenge for this step is to find co-references across large text spans, because for our application these cannot be limited to references within sentences or paragraphs.

Concept labeling: Once the set of equivalent concepts is produced by the previous step, they are assigned a unique label. The input context concept map is used for this purpose. All concept labels from the context concept map are extracted and compared with the sets of normalized concepts, using the procedure described in the previous step. If there is a match, the set of concepts is assigned the label from the context concept map. Otherwise, it is assigned the shortest label extracted from the document. For example, the normalized concept set: {“line of thunderstorms”, “thunderstorm activity”} is labeled as “thunderstorms”, instead of “thunderstorm activity” making it more similar to the context concept map, therefore augmenting the chances of being classified in this category.

Linking phrase extraction: Using the parsed sentences and normalized concepts, the sentences in the document 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. For example, “thunderstorms” –are frequent in→ “the gulf coast”.

Concept map generation: The information from the extracted concepts and linking phrases, in the form of propositions, is used to construct a graphical representation of the concept map. Although this 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. Finally, after integration of all propositions, the map can contain node strings (sequences of nodes that are not connected to other segments) and these are replaced by a single node whose label is the concatenation of the node string labels. This replacement has minor effects on the individual node weight during concept map index comparison.

Figure 3 shows an example of concept map indices generated from a document by the system. The top concept map is an input context map used as context for index generation. The bottom left map is an index concept map generated by the previous version of the algorithm without the context-based concept-labeling step, and the bottom right map is the index generated by the new algorithm. The highlighted concepts correspond to concepts that were matched during the labeling step and were replaced. The document passage from which the indices were generated is shown at the bottom of the figure.



Fronts
are defined in terms of
are the
have
types are
Temperature and moisture differences
Boundary
Developmental phases
Warm
Occluded
Stationary fronts
Cold fronts
between
projected as
which are
are
are more common in
Interacting air masses
Lines on a surface chart
Frontogenesis (strengthening)
Frontolysis (dissipation)
Rare in the Gulf Coast region
The Gulf Coast region
due to
because of
due to the local influence of
Tendency of Low pressure centers to migrate across CONUS to the north of the region
Migratory tendencies of the Polar jet
Maritime air mass
Several different types of air masses
affect
The continental United States
The Gulf Region
is particularly influenced by
is influenced by
The maritime tropical air mass
The position of the jet
form in
is produced over
is
The tropics or subtropics over the low latitudes
oceans
warm and humid
Several different types of air masses
affect
The continental United States
The Gulf Coast region
is particularly influenced by
is influenced by
Maritime air mass
The position of the jet
form in
is produced over
is
The tropics or subtropics over the low latitudes
oceans
warm

Several different types of air masses affect the continental United States. The Gulf region is particularly influenced by the maritime Tropical air mass in the summer when the jet stram has migrated north. It is influenced by the position of the jet and the interplay of the continental Polar and maritime Tropical air masses in the winter. Maritime Tropical air masses form in the tropics or subtropics over the low latitudes. Maritime tropical air is produced over oceans and is warm and humid.

Figure 3. Example of a document converted to a concept map (top map is from STORM-LK (Hoffman *et al.* 2001)).

6 Concept Map Similarity Assessment

To identify relevant concept maps, the index concept maps are compared with the corresponding context concept map using cosine similarity (Baeza-Yates & Ribeiro-Neto 1999) and a vector-model representation of concept maps (Leake *et al*. 2003). The concept map vectors are constructed as in (Valerio, Leake, & Cañas 2007), using the Hub-Authority-Root-Distance (HARD) model (Reichherzer & Leake 2006b) to estimate concept importance based on structural features, each concept is assigned a weight based on its *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). 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.

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 (Cañas, Leake, & Maguitman 2001).

In our experiments, the parameters are set to $\phi = 0, \psi = 2.235, \gamma = 1.764$, which were previously found to best fit the model for experimental user data (Leake, Maguitman & Reichherzer 2004). The weight *w*(*j*) of keyword *j* is the sum of the concept weights multiplied by the frequency of the keyword in each concept.

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

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

6.1 Experimental setup

Our experiment tests the ability of the algorithm to associate an input document to the most relevant maps in a collection of concept maps constructed by experts. The test data for the evaluation is a set of existing knowledge models containing a number of concept maps annotated with topically related documents, which have been used previously as "gold standard" concept maps for evaluating concept map-document associations. The knowledge models from Mars 2001 (Briggs *et al*. 2004) and STORM-LK (Hoffman *et al*. 2001) contain a total of 80 concept maps and 131 different documents already linked to the concept maps. It is possible for a document to be associated with more than one concept map. The evaluation is based on a match between the concept maps identified by the system as the most relevant and the original concept map annotations, measuring the ability of the procedure to find the original associations.

To perform the test, all documents are separated from the concept maps. Next, each of the documents is processed individually with no prior knowledge about the concept maps to which it was originally linked used in this processing. As described in the previous section, for each document the concept map generation process is repeated with all 80 concept maps separately, producing 80 slightly different concept map indices differing on their concept labels. The system then compares the produced index concept maps with the corresponding concept maps in the knowledge model using the similarity measure describe above. Next, the concept map indices are sorted in descending order by their similarity value to the maps used as context for generating them, with the similarity measure used to judge relevance.

One goal of this evaluation is to determine the precision and recall achieved by the system when different cutoffs are applied to select the relevant concept maps from the sorted list. In our case, the cutoffs range from 1 to 5. Cutoff = 2 means that the two most similar concept maps are attached to the document. An attachment is considered successful if the document is correctly associated with a concept map originally containing it.

6.2 Experimental results

The algorithm performance was compared to the algorithm presented in (Valerio, Leake, & Cañas 2007) and to a baseline algorithm that constructs its document vector representation solely based on keyword frequency. The latter illustrates the performance in the absence of structural information.

Figure 4 shows the results of the evaluation. The new algorithm showed an average precision increase of 14% compared to the previous algorithm that does not use the target concept map labels, and 27% compared to the baseline. We also calculated the F1 measure (harmonic mean of precision/recall) when only the most similar concept is associated with the document (cutoff = 1). In this case, the proposed algorithm also outperformed the other methods by similar margins. This indicates that the precision was increased without degrading recall.

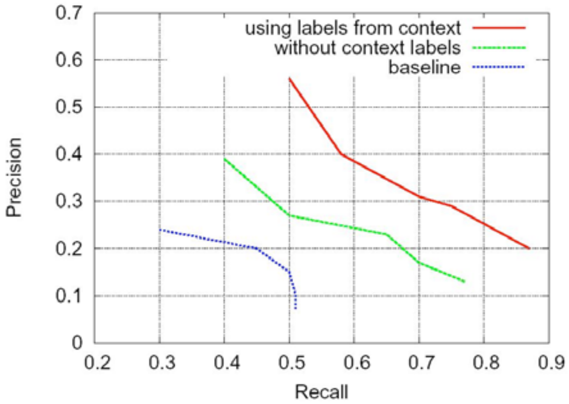


Figure 4. Precision/Recall plot for document classification with the three methods.

The improvement when a concept map index is constructed using the concept labels from a target concept map suggests the value of using the context of the target concept maps to refine the automatic concept map generation procedure, indicating that the information obtained from the concept map context is meaningful. These results also indicate a significant improvement of the results compared to the keyword-based algorithm reaffirming that the structure of the generated concept map gives valuable information during the document classification task.

7 Summary and future work

This paper presented a top-down/bottom-up algorithm to extract information from documents to construct concept map indices automatically, using target concept maps as context to refine the assignment of concept labels. The addition of top-down information resulted in a significant performance improvement compared to the previous bottom-up only approach, when using the indices for a document classification task. These results suggest the promise of this approach to generating concept-map indices from documents, taking advantage of existing natural language processing techniques to extract information efficiently from documents and at the same time using existing concept map knowledge models to guide the construction process as a higher-level semantic information source.

The ultimate goal of our project is to develop intelligent user interfaces to assist during document understanding and contextualization tasks. For this work, we intend to further refine the concept normalization step of the conversion procedure to produce better quality concept map indices and to also refine and evaluate the linking phrase extraction step, which we foresee as an interesting and challenging task.

References

Abney, S. P. (1996). Part-of-Speech Tagging and Partial Parsing. In Church, K., Young, S., and Bloothooft, G., (Eds.), *Corpus-Based Methods in Language and Speech*. Kluwer Academic Publishers.

Alves, A. O., Pereira, F. C. & Cardoso, A. (2001). Automatic Reading and Learning from Text. In Proceedings of the International Symposium on Artificial Intelligence (ISAI-2001), pp. 302–310.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). Modern Information Retrieval. ACM Press/Addison-Wesley.

Briggs, G., Shamma, D. A., Cañas, A. J., Carff, R., Scargle, J., & Novak, J. D. (2004). Concept Maps Applied to Mars Exploration Public Outreach. In A. J. Cañas, J. D. Novak & F. González (Eds.), *Concept Maps: Theory, Methodology, Technology. Proceedings of the First International Conference on Concept Mapping* (Vol. I, pp. 109-116). Pamplona, Spain: Universidad Pública de Navarra.200

Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Eskridge, T. C.; Arroyo, M.; and 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* (Vol. I, pp. 125-133). Pamplona, Spain: Universidad Pública de Navarra.

Cañas, A. J.; Leake, D. B.; and Maguitman, A. G. (2001). Combining Concept Mapping with CBR: Towards Experience-Based Support for Knowledge Modeling. In *Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference*, pp. 286–290. AAAI Press.

Clariana, R. B., & Koul, R. (2004). A Computer-Based Approach for Translating Text into Concept Map-like Representations. 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* (Vol. I). Pamplona, Spain: Universidad Pública de Navarra.

Fellbaum, C., ed. (1998). WordNet: An Electronic Lexical Database. MIT Press.

Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Hickl, A. & Wang, P. (2005). Employing Two Question Answering Systems in TREC-2005. In *Proceedings of the 14th Text Retrieval Conference* (TREC 2005).

Hoffman, R. R., Coffey, J. W., Ford, K. M. & Carnot, M. J. (2001). STORM-LK: A Human-Centered Knowledge Model For Weather Forecasting. In *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society*.

Leake, D. B., Maguitman, A., Reichherzer, T., Cañas, A. J., Carvalho, M., Arguedas, M., Brenes, S., and Eskridge, T. (2003). Aiding Knowledge Capture by Searching for Extensions of Knowledge Models. In *Proceedings of the Second International Conference on Knowledge Capture* (K-Cap 2003), pp. 44–53.

Leake, D. B., Maguitman, A., Reichherzer, T., Cañas, A. J., Carvalho, M., Arguedas, M., and Eskridge, T. C. (2004). Googling from a Concept Map: Towards Automatic Concept-Map-Based Query Formation. 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* (Vol. I, pp. 409-416). Pamplona, Spain: Universidad Pública de Navarra.

Leake, D. B., Maguitman, A., & Reichherzer, T. (2004). Understanding Knowledge Models: Modeling Assessment of Concept Importance in Concept Maps. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 795-800). Mahwah, NJ: Lawrence Erlbaum.

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

Rajaraman, K., & Tan, A.-H. (2002). Knowledge Discovery from Texts: A Concept Frame Graph Approach. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, pp. 669–671.

Reichherzer, T., & Leake, D. B. (2006a). Towards Automatic Support for Augmenting Concept Maps with Documents. In A. J. Cañas & J. D. Novak (Eds.), *Concept Maps: Theory, Methodology, Technology. Proceedings of the Second International Conference on Concept Mapping* (Vol. 1). San Jose, Costa Rica: Universidad de Costa Rica.

Reichherzer, T., & Leake, D. B.. (2006b). Understanding the Role of Structure in Concept Maps. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*, 2004–2009.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys 34(1):1–47.

Valerio, A., & Leake, D. B. (2006). Jump-Starting Concept Map Construction with Knowledge Extracted From Documents. In A. J. Cañas & J. D. Novak (Eds.), *Concept Maps: Theory, Methodology, Technology. Proceedings of the Second International Conference on Concept Mapping*. San Jose, Costa Rica: Universidad de Costa Rica.

Valerio, A., Leake, D., & Cañas, A. J. (2007). Automatically Associating Documents with Concept Map Knowledge Models. In *Proceedings of the Thirty-third Latin American Conference in Informatics (CLEI 2007),* San José, Costa Rica, Oct 2007.