Evaluating Concept Maps As A Cross-Language Knowledge Discovery Tool for NDLTD

Ryan Richardson
Ph.D. Candidate, Virginia Tech, Blacksburg, Virginia

Edward A. Fox
Professor of Computer Science, Virginia Tech

John Woods
Undergraduate Student, Virginia Tech

Keywords: Concept maps, automatic summarization, phrase translation, NDLTD

ABSTRACT

Concept maps, first suggested by Joseph Novak, have been extensively studied in the education field as an aid for learners to increase understanding. We hypothesize that concept maps could be helpful summaries for large documents (like ETDs), including those in other languages, since they contain the most important concepts, which can be translated easily, without regard to creating natural flowing text (e.g., abstracts). They also show relations between concepts, which keyword lists cannot. Hence we believe concept maps could allow researchers to discover pertinent dissertations in languages they cannot read, helping them to decide if they want a potentially relevant dissertation translated.

We are evaluating concept maps as summaries for electronic dissertations, comparing them with abstracts and keywords. We have conducted two user studies, one with expert-drawn concept maps and the other with automatically generated concept maps, and plan more user studies. We will evaluate both expert-drawn concept maps and ones automatically generated using term co-occurrence, for both English and Spanish dissertations. To translate the concept maps, we will compare specialized bilingual phrase extraction tools based on work done at the Universidad Nacional de Educación a Distancia (UNED) in Madrid to expert human translations. The concept maps will be evaluated against machine and expert translations of abstracts and/or keywords. Our test collection has approximately 50 computer science ETDs from Virginia Tech and 100 from Universidad de las Américas, in Puebla, Mexico.

1. INTRODUCTION: THE CONCEPT MAP

Since the 1980's researchers in the education field have studied concept maps as a means to facilitate the quick and effective learning of concepts (Coffey et al., 2003). Concepts are defined as "perceived regularities in events or objects, or records of events of objects, designated by a label" (Novak, 1998). Concept maps (Novak & Gowin, 1984) can be defined as "graphical representations of knowledge that are comprised of concepts and the relationships between them". They consist of labeled nodes and links and are regarded as useful pedagogical tools. As envisaged by Novak, in a concept map concepts are presented in an hierarchical fashion with the most general concepts at the top, and with more specific, less general concepts arranged below. Figure 1 below illustrates this approach to concept maps, and introduces many of the key themes discussed in the remainder of this document.

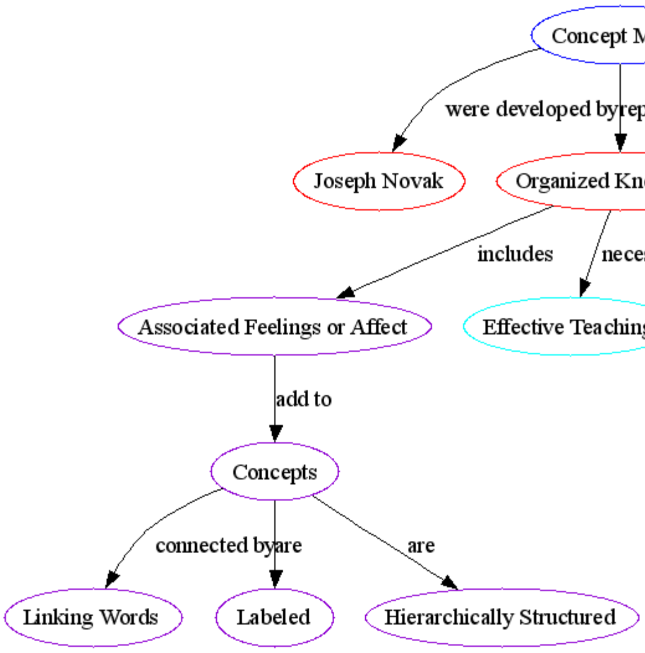


Figure 1: Example Concept Map

There has been much research on how concept maps allow students to acquire knowledge more quickly than traditional teaching techniques (McNaught & Kennedy, 1997). However, there has been limited research on using concept maps as summaries for a paper or group of papers. Currently the two most common summaries of research papers are abstracts and keyword lists. Keyword lists are inadequate because they cannot show relationships between sets of keywords. Suppose a researcher is looking for a document where topic A relates to topic B with relationship C. The appearance of A and B in the keyword list does not necessarily mean that the document is relevant to the researcher. Further, the problems with using abstracts as the primary summarization of a document are myriad (Levin & Redell, 1983). Writing good abstracts is difficult, and often the abstract mentions some topics which are only briefly discussed in the paper, and omits other topics which make up large sections of the paper. Many researchers, while technically proficient, consistently follow poor sentence organization practices, creating virtually-unreadable abstracts (Gopan & Swan, 1990) . Finally, the length and form of abstracts vary widely across subject areas and especially across languages. When it comes to automatically translating abstracts, the difficulties become even more obvious. Anyone who has ever used BabelFish (tm) to translate an abstract into their preferred language knows the poor readability and ambiguities in the resulting text.

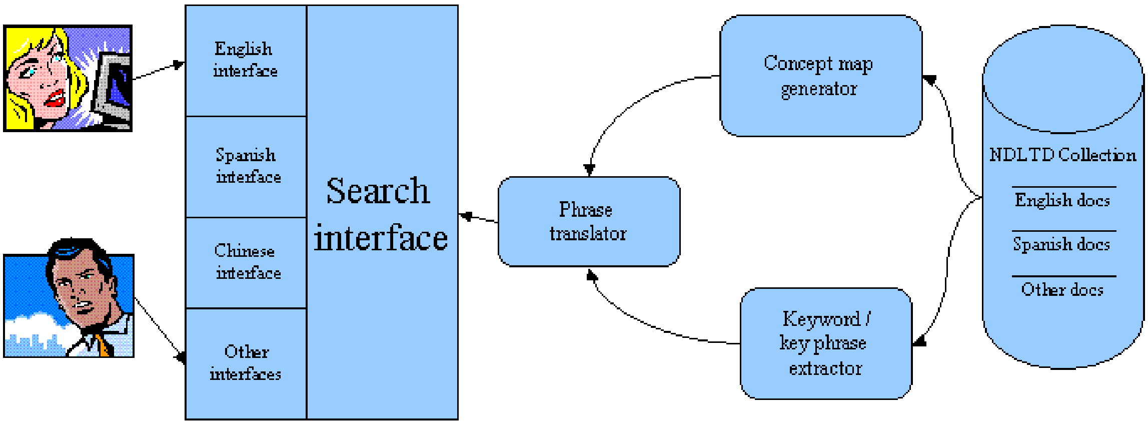


Figure 2: Proposed solution to NDLTD cross-language problem

2. THE NDLTD CROSS-LANGUAGE PROBLEM

The Networked Digital Library of Theses and Dissertations (NDLTD, 2005) contains metadata for about 200,000 (ETDs). Table 1 shows a language breakdown for part of the collection, as of July 2005 (Saskia van Acker, personal communication, July 12, 2005).

Table 1: Language Breakdown for NDLTD

Language	**Number**
English	123,696
Portuguese	11434
German	4131
French	3868
Spanish	1561
Chinese	1463
Catalan	804
Swedish	348
Dutch	21
Norwegian	6
Danish	6
Finnish	2
Unclassified	19579
Total	**166919**

Thus there now are well over 23,000 non-English theses and dissertations that English-speakers might wish to read. As more universities around the globe participate in NDLTD, there will be tens of thousands of additional works. Further, those who prefer to use languages other than English would appreciate access to English ETDs from their native language. Given the size of the average dissertation, however, researchers might have trouble determining if one of these dissertations is relevant to their research. Our hypothesis is that automatically generated and translated concept maps could aid researchers in making relevance judgments for ETDs.

3. FEASIBILITY OF USING CONCEPT MAPS AS A CROSS-LANGUAGE TOOL

To test the hypothesis that concept maps can be a better summary than an abstract, both for one language and across languages, we conducted an initial experiment. We used concept maps drawn by experts, to see if students could identify relevant papers. The concept maps were based on the full text of the papers, so they might contain information not in the abstract. The experiment had a mono-lingual (English-only) component and a cross-lingual (Spanish to English) component. Users were given four relevance tasks (two for mono-lingual and two for cross-lingual). For all four tasks users performed better when having concept maps in addition to abstracts. The differences were only significant in 1 of the 4 cases however. Users also reported that having the concept map available in their language was significantly more helpful than just having the abstract (Richardson & Fox, 2005).

4. AUTOMATIC GENERATION OF CONCEPT MAPS

We have experimented with automatically producing concept maps using associations between words in the documents. We evaluated several techniques to find the associations, including t-scores, association rules (Agrawal, Imielinski, & Swami, 1993), mutual information (Hindle, 1990), and the Dice co-efficient (Frakes & Baeza-Yates, 1992). Below is an example of some phrases extracted from the thesis “How Politics and Culture Affected the Designs of the U.S. Space Shuttle and the Soviet Buran” by Stephen Garber.

Table 2: **Top eight relations from a masters thesis, using association rules, t-scores, mutual information, and Dice coefficient.**

Association rules	**t-score**	**Mutual Information**	**Dice coefficient**
NASA **<>** transportation	Energiya**<>**buran	interview**<>** stephen_garber	military_characterization**<>** reluctant_support
NASA **<>** option	interview**<>** stephen_garber	range**<>**capability	military_characterization**<>** military
Shuttle **<>** U.S._space	Range**<>**capability	johnson**<>**soviet_year	joseph_P**<>**hopkin
buran <- energiya	NASA**<>**shuttle	energiya**<>**buran	blueprint**<>**misleading
soviet **<>** space_technology	Cross**<>**range	cross**<>**range	myriad_american**<>**blueprint
soviet **<>** aviation_week	Stephen_garber**<>** November	izvestiya**<>**december	soviet_figures**<>**chronology
space **<>** station	Johnson**<>**soviet_year	flying**<>**december	julian_E**<>**settings
space **<>** transportation	interview**<>**november	yaroslav_golvanov**<>** december	key**<>**soviet_bibliography

As one can see from Table 2, the associations found by t-scores and association rules convey more general topics found in the paper. Mutual information and Dice coefficient tend to find fixed phrases that occurred rarely in the paper. We chose to concentrate on association rules for our automatic concept map generation.

We experimented with three design layouts for concept maps. The first is a whole document map, which draws a single concept map of the entire document. The second is a chapter-level map, which uses the same algorithm to draw a map of each chapter. The third type is a table-of-contents map, which uses the dissertation's own table of contents as a skeleton of the map, and then embellishes it with additions found with association rules.

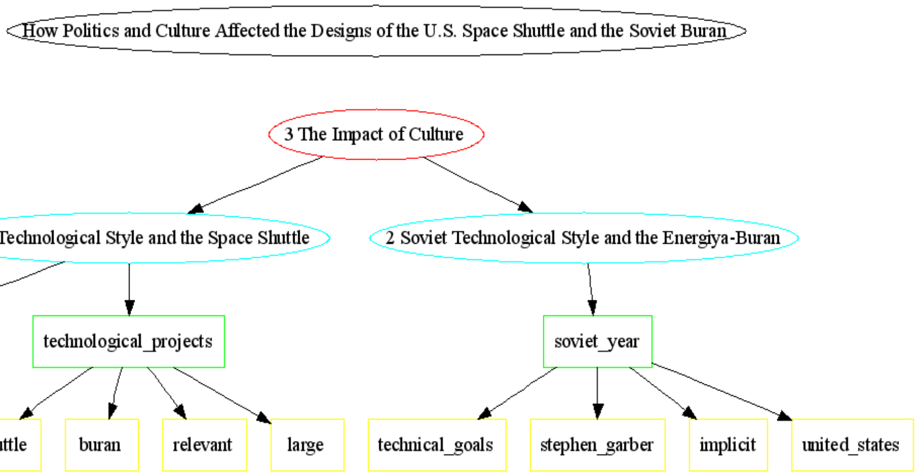
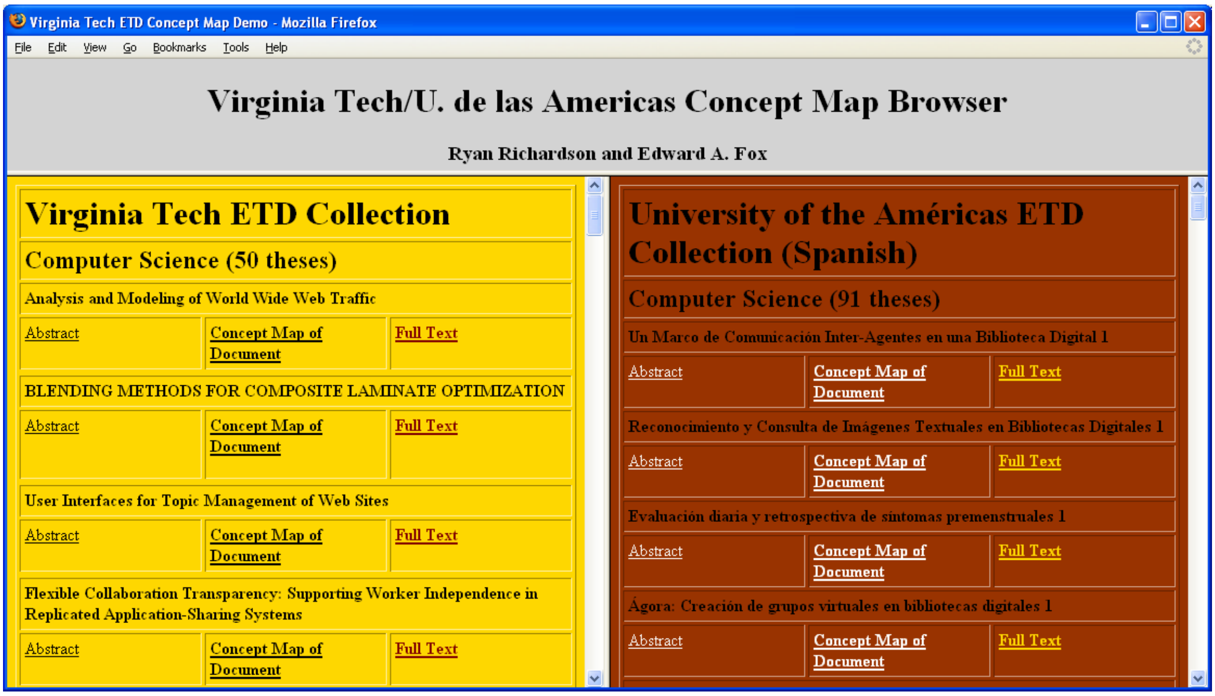

How Politics and Culture Affected the Designs of the U.S. Space Shuttle and the Soviet Buran
3 The Impact of Culture
1 U.S. Technological Style and the Space Shuttle
2 Soviet Technological Style and the Energiya-Buran
space
technological_projects
soviet_year
N1
johnson
U.S._shuttle
buran
relevant
large
technical_goals
stephen_garber
implicit
united_states

Figure 3: Table of contents map of thesis "How Politics and Culture Affected the Design of the U.S. Space Shuttle and the Soviet Buran", chapter 3.

We conducted an experiment to determine which of the three designs users preferred. Users significantly preferred the table-of-contents layout over the other two. Details can be found in (Richardson & Fox, 2005).

We automatically generated concept maps for 50 ETDs from Virginia Tech and 91 theses from Universidad de Las Américas in Puebla, Mexico. The English ETDs have concepts maps in both the chapter-level and table-of-contents formats. The Spanish ETDs have concept maps in the chapter-level format (table-of-contents information was not available for these ETDs). Our website shows these concept maps, in addition to abstracts and full text (http://www.dlib.vt.edu/~ryanr/cm_demo) .



Virginia Tech/U. de las Americas Concept Map Browser

Ryan Richardson and Edward A. Fox

Virginia Tech ETD Collection

Computer Science (50 theses)

Analysis and Modeling of World Wide Web Traffic

Abstract | Concept Map of Document | Full Text

BLENDING METHODS FOR COMPOSITE LAMINATE OPTIMIZATION

Abstract | Concept Map of Document | Full Text

User Interfaces for Topic Management of Web Sites

Abstract | Concept Map of Document | Full Text

Flexible Collaboration Transparency: Supporting Worker Independence in Replicated Application-Sharing Systems

Abstract | Concept Map of Document | Full Text

University of the Américas ETD Collection (Spanish)

Computer Science (91 theses)

Un Marco de Comunicación Inter-Agentes en una Biblioteca Digital 1

Abstract | Concept Map of Document | Full Text

Reconocimiento y Consulta de Imágenes Textuales en Bibliotecas Digitales 1

Abstract | Concept Map of Document | Full Text

Evaluación diaria y retrospectiva de síntomas premenstruales 1

Abstract | Concept Map of Document | Full Text

Ágora: Creación de grupos virtuales en bibliotecas digitales 1

Abstract | Concept Map of Document | Full Text

Figure 4: ETD concept map browser interface

5. AUTOMATIC TRANSLATION OF CONCEPT MAPS

The next step for using automatically generated concept maps is to be able to automatically translate them well enough so that an end-user will find them useful. Since concept maps need not have natural flowing text like an abstract, we feel this is a more realistic goal than translating abstracts. Toward this end we are currently implementing a phrase-extraction and translation tool along the lines of (López-Ostenero, Gonzalo, & Verdejo, 2004), using a statistical test found in (Evans & Zhai, 1996). It uses an English and Spanish comparable corpus which was derived from ETD collections at Virginia Tech and UDLA. It also uses a lexicon developed at the University of Maryland, to create a Markov model of phrases for English and Spanish. To test the phrase translation, we plan to use the collection of 20,000 concept maps at the Institute of Human and Machine Cognition (IHMC) (Cañas, 2005) at http://ihmc.us/. This collection consists of concept maps in English, Spanish, Portuguese, and Italian.

6. CONCLUSION

Due to the proliferation of ETDs in multiple languages available via NDLTD, the problem of finding useful information in dissertations that one cannot read will only become more important over time. We have preliminary results indicating that concept maps could be a useful summarization for ETDs, both in the mono-lingual and cross-lingual case. We have tested three styles of automatically generated concept maps. Now we are concentrating on the next step of the process, which is automatically translating concept maps, first on hand-drawn concept maps and later on automatically-generated ones.

7. ACKNOWLEDGMENTS

This work was partially supported by NSF grants DUE-0121679 and IIS-0080748, and by SAIC as part of the Research and Development Experimental Collaboration—Applied

Research Center. We thank Fernando Lopez-Osteñero for the method used in our translations, and Doug Oard and Daqing He at U. of Maryland for their Spanish lexicon.

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993, May). *Mining association rules between sets of items in large databases.* Paper presented at the ACM SIGMOD Conference on Management of Data, Washington, D.C.

Cañas, A. J. (2005). *Institute for Human and Machine Cognition*, 2005, from http://ihmc.us/

Coffey, J. W., Carnot, M. J., Feltovich, P. J., Hoffman, R. R., Cañas, A. J., & Novak, J. D. (2003). *A Summary of Literature Pertaining to the Use of Concept Mapping Techniques and Technologies for Education and Performance Support*. Pensacola, FL: Chief of Naval Education and Training.

Evans, D. A., & Zhai, C. (1996, June 24-27). *Noun-Phrase Analysis in Unrestricted Text for Information Retrieval.* Paper presented at the Association for Computational Linguistics, Santa Cruz, California.

Frakes, W. B., & Baeza-Yates, R. (1992). *Information Retrieval, Data Structure and Algorithms*: Prentice Hall.

Gopan, G. D., & Swan, J. A. (1990). The science of scientific writing. *American Scientist, 78*, 550-558.

Hindle, D. (1990). *Noun classification from predicate-argument structures.* Paper presented at the ACL-90, Pittsburgh.

Levin, R., & Redell, D. D. (1983). An Evaluation of the Ninth SOSP Submissions or How and How Not to Write a Good Systems Paper. *ACM SIGOPS Operating Systems Review, 17*(3), 35-40.

López-Ostenero, F., Gonzalo, J., & Verdejo, F. (2004). Noun Phrases as building blocks for Cross-Language Search Assistance. *Information Process and Management, 41*, 549-568.

McNaught, C., & Kennedy, D. (1997). Use of Concept Mapping in the design of learning tools for interactive multimedia. *Journal of Interactive Learning Research, 8*(3-4), 389-406.

NDLTD. (2005). Networked Digital Library of Theses and Dissertations. http://www.ndltd.org.

Novak, J. D. (1998). *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in schools and Corporations.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Novak, J. D., & Gowin, D. B. (1984). *Learning How To Learn*. Cambridge, UK: Cambridge University Press.

Richardson, R., & Fox, E. A. (2005, June 7-11). *Using concept maps in digital libraries as a cross-language resource discovery tool.*

Paper presented at the Joint Conference on Digital Libraries (JCDL 2005), Denver.