Pedagogically Useful Extractive Summaries for Science Education

Sebastian de la Chica, Faisal Ahmad, James H. Martin, Tamara Sumner
Institute of Cognitive Science
Department of Computer Science
University of Colorado at Boulder
sebastian.delachica, faisal.ahmad, james.martin,
tamara.sumner@colorado.edu

Abstract

This paper describes the design and evaluation of an extractive summarizer for educational science content called COGENT. COGENT extends MEAD based on strategies elicited from an empirical study with science domain and instructional design experts. COGENT identifies sentences containing pedagogically relevant concepts for a specific science domain. The algorithms pursue a hybrid approach integrating both domain independent bottom-up sentence scoring features and domain-aware top-down features. Evaluation results indicate that COGENT outperforms existing summarizers and generates summaries that closely resemble those generated by human experts. COGENT concept inventories appear to also support the computational identification of student misconceptions about earthquakes and plate tectonics.

1 Introduction

Multidocument summarization (MDS) research efforts have resulted in significant advancements in algorithm and system design (Mani, 2001). Many of these efforts have focused on summarizing news articles, but not significantly explored the research issues arising from processing educational content to support pedagogical applications. This paper describes our research into the application of MDS techniques to educational science content to generate pedagogically useful summaries.

Knowledge maps are graphical representations of domain information laid out as networks of nodes containing rich concept descriptions interconnected using a fixed set of relationship types (Holley and Dansereau, 1984). Knowledge maps are a variant of the concept maps used to capture, assess, and track student knowledge in education research (Novak and Gowin, 1984). Learning research indicates that knowledge maps may be useful cognitive scaffolds, helping users lacking domain expertise to understand the macro-level structure of an information space (O'Donnell et al., 2002). Knowledge maps have emerged as an effective representation to generate conceptual browsers that help students navigate educational digital libraries, such as the Digital Library for Earth System Education (DLESE.org) (Butcher et al., 2006). In addition, knowledge maps have proven useful for domain and instructional experts to capture domain knowledge from digital library resources and to analyze student understanding for the purposes of providing formative assessments (Ahmad et al., 2007).

Knowledge maps have proven useful both as representations of knowledge for assessment purposes and as learning resources for presentation to students. However, domain knowledge map construction by experts is an expensive knowledge engineering activity. In this paper, we describe our progress towards the automated generation of pedagogically useful extractive summaries from educational texts about a science domain. In the context of automated knowledge map generation, summary sentences correspond to concepts. While the detection of relationships between concepts is also part of our overall research agenda, this paper focuses solely on concept identification using MDS techniques. The remainder of this paper is organized as fol-

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* license (http://creativecommons.org/licenses/by-nc-sa/3.0/). Some rights reserved.

lows. First, we review related work in the areas of automated concept extraction from texts and extractive summarization. We then describe the empirical study we have conducted to understand how domain and instructional design experts identify pedagogically important science concepts in educational digital library resources. Next, we provide a detailed description of the algorithms we have designed based on expert strategies elicited from our empirical study. We then present and discuss our evaluation results using automated summarization metrics and human judgments. Finally, we present our conclusions and future work in this area.

2 Related Work

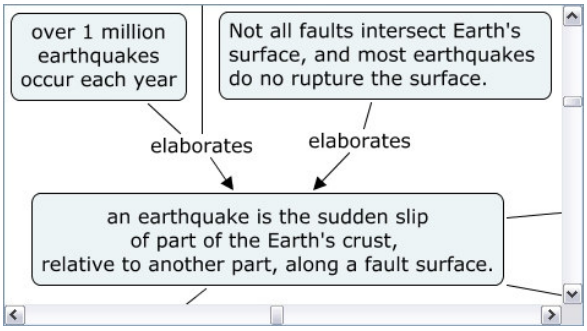
Our work is informed by efforts to automate the acquisition of ontology concepts from text. OntoLearn (Navigli and Velardi, 2004) extracts domain terminology from a collection of texts using a syntactic parse to identify candidate terms that are filtered based on domain relevance and connected using a semantic interpretation based on word sense disambiguation. The newly identified concepts and relationships are used to update an existing ontology. Knowledge Puzzle focuses on n-grams to produce candidate terms filtered based on term frequency in the input documents and on the number of relationships associated with a given term (Zouaq et al., 2007). This approach leverages pattern extraction techniques to identify concepts and relationships. While these approaches produce ontologies useful for computational purposes, the identified concepts are of a very fine granularity and therefore may yield graphs not suitable for identifying student misconceptions or for presentation back to the student. Clustering by committee has also been used to discover concepts from a text by grouping terms into conceptually related clusters (Lin and Pantel, 2002). The resulting clusters appear to be tightly related, but operate at a very fine level of granularity. Our approach focuses on sentences as units of knowledge to produce concise representations that may be useful both as computational objects and as learning resources to present back to the student. Therefore, extractive summarization research also informs our work.

Topic representation and topic themes have been used to explore promising MDS techniques (Harabagiu and Lacatusu, 2005). Recent efforts in graph-based MDS have integrated sentence affinity, information richness and diversity penalties to produce very promising results (Wan and Yang, 2006). Finally, MEAD is a widely used multi-document summarization and evaluation platform (Radev et al., 2000). MEAD research efforts have resulted in significant contributions to support the development of summarization applications (Radev et al., 2000). While all these systems have produced promising results in automated evaluations, none have directly targeted educational content as input or the generation of pedagogically useful summaries. We are directly building upon MEAD due its focus on sentence extraction and its high degree of modularization.

3 Empirical Study

We have conducted a study to capture how human experts construct and use knowledge maps. In this 10-month study, we examined how experts created knowledge maps from educational digital libraries and how they used the maps to assess student work and provide personalized feedback.

In this paper, we are focusing on the knowledge map construction aspects of the study. Four geology and instructional design experts collaboratively selected 20 resources from DLESE to construct a domain knowledge map on earthquakes and plates tectonics for high school age learners. The experts independently created knowledge maps of individual resources which they collaboratively merged into the final domain knowledge map in a one-day workshop. The resulting domain knowledge map consisted of 564 nodes containing domain concepts and 578 relationships. The concepts consist of 7,846 words, or 5% of the total number of words in the original resources. Figure 1 shows a fragment of the domain knowledge map created by our experts.



over 1 million earthquakes occur each year

Not all faults intersect Earth's surface, and most earthquakes do no rupture the surface.

an earthquake is the sudden slip of part of the Earth's crust, relative to another part, along a fault surface.

Figure 1. Fragment of domain knowledge map created by domain and instructional experts

Experts created nodes containing concepts of varying granularity, including nouns, noun phrases, partial sentences, single sentences, and multiple sentences. Our analysis of this domain knowledge map indicates that experts relied on copying-and-pasting (58%) and paraphrasing (37%) to create most domain concepts. Only 5% of the nodes could not be traced directly to the original resources.

Experts used relationship types in a Zipf-like distribution with the top 10 relationship types accounting for 64% of all relationships. The top 2 relationship types each accounted for more than 10% of all relationships: elaborations (19% or 110 links) and examples (14% or 78 links).

We have established the completeness of this domain knowledge map by asking a domain expert to assess its content coverage of nationally-recognized educational goals on earthquakes and plate tectonics for high school age learners using the American Association for the Advancement of Science (AAAS) Benchmarks (Project 2061, 1993). The results indicate adequate content coverage of the relevant *AAAS Benchmarks* achieved through 82 of the concepts (15%) with the remaining 482 concepts (85%) providing very detailed elaborations of the associated learning goals.

Qualitative analysis of the verbal protocols captured during the study indicates that all experts used external sources to construct the domain knowledge map. Experts made references to their own knowledge (e.g., "I know that…"), to content learned or taught in geology courses, to other resources used in the study, and to the National Science Education Standards (NSES), a comprehensive collection of nationally-recognized science learning goals for K-12 students (National Research Council, 1996).

We have examined sentence extraction agreement between experts using a kappa measure that accounts for prevalence of judgments and conflicting biases amongst experts, called PABA-kappa (Byrt et al., 1993). The average PABA-kappa value of 0.62 indicates that our experts substantially agree on sentence extraction from digital library resources. While this study was not designed as an annotation project to support summarization evaluation, this level of agreement indicates that the concepts selected by the experts may serve as the reference summary to evaluate the performance of our summarizer.

4 Summarizer for Science Education

Creating a knowledge map from a collection of input texts involves identifying sentences containing important domain concepts, linking concepts, and labeling those links. This paper focuses solely on identifying and extracting pedagogically relevant sentences as domain concepts. We have designed and implemented an extractive summarizer for educational science content, called COGENT, based on MEAD version 3.11 (Radev et al., 2000). COGENT processes a collection of educational digital library resources by first preprocessing each resource using Tidy (tidy.sourceforge.net) to fix improperly formatted HTML code. COGENT then merges multiple web pages into a single HTML document and extracts the contents of each resource into a plain text file. We have extended MEAD with sentence scoring features based on domain content, document structure, and sentence length.

4.1 Domain Content

We have designed two sentence-scoring features that aim to capture the domain content relevance of each sentence: the educational standards feature and the gazetteer feature.

We have developed a feature that models how human experts used external sources to identify and extract concepts. The *educational standards* feature uses the textual description of the relevant AAAS Benchmarks on earthquakes and plate tectonics for high-school age learners and the associated NSES. Each sentence receives a score based on its similarity to the text contents of the learning goals and educational standards computed using a TFIDF (Term Frequency-Inverse Document Frequency) approach (Salton and Buckley, 1988). We have used KinoSearch, a Perl implementation of the Lucene search engine (lucene.apache.org), to create an index that includes the AAAS Benchmarks learning goal description (boosted by 2), subject (boosted by 8), and keywords (boosted by 2), plus the text of the associated national standards (not boosted). Sentence scores are based on the similarity scores generated by KinoSearch in response to a query consisting of the sentence text.

To account for the large number of examples used by the experts in the domain knowledge map (14% of all links), we have developed a feature that reflects the number and relevance of the geographical names in each sentence. Earth science examples often refer to names of geographical places, including geological formations on the planet. The *gazetteer* feature leverages the Alexandria Digital Library (ADL) Gazetteer service (Hill, 2000) to check whether named entities identified in each sentence match entries in the ADL Gazetteer. A gazetteer is a georeferencing resource containing information about locations and place-names, including latitude and longitude as well as type information about the corresponding geographical feature. Each sentence receives a score based on a TFIDF approach where the TF is the number of times a particular location name appears in the sentence and the IDF is the inverse of the count of gazetteer entries matching the location name. If the ADL Gazetteer returns a large number of results for a given place-name, it means there are many geographical locations identified by that name. Our assumption is that unique names may be more pedagogically relevant. For example, Ohio receives an IDF score of 0.0625 because the ADL Gazetteer contains 16 entries so named, while the Mid-Atlantic Ridge, the distinctive underwater mountain range dividing the Atlantic Ocean, receives a score of 1.0 as it appears only once.

4.2 Document Structure

Based on the intuition that the HTML structure of a web site reflects content relevancy, we have developed the hypertext feature. The *hypertext* feature assigns a higher score to sentences contained under higher level HTML headings.

Heading	Bonus
H1	1/1 = 1.00
H2	1/2 = 0.50
H3	1/3 = 0.33
H4	1/4 = 0.25
H5	1/5 = 0.20
H6	1/6 = 0.17

Table 1. Hypertext feature heading bonus

Within a given heading level, the hypertext feature assigns a higher score to sentences that appear earlier within that level based on both relative paragraph order within the heading and relative sentence position within each paragraph. The equation used to compute the hypertext score for a sentence is

$$hypertext\_score = heading\_bonus * \sqrt[4]{1/par\_no} * \sqrt[4]{1/sent\_no}$$

where *heading_bonus* is obtained from Table 1, *par_no* is the paragraph number within the heading, and *sent_no* is the sentence number within the paragraph. We use the $\sqrt[4]{1/x}$ function to attenuate the contributions to the feature score of later paragraphs and sentences. Initially, we used the same function MEAD uses to modulate its position feature ($\sqrt[2]{1/x}$), but initial experimentation indicated this function decayed too rapidly, resulting in later sentences being over-penalized.

4.3 Sentence Length

To promote the extraction of sentences containing scientific concepts, we have developed the *content word density* feature. This feature makes a cut-off decision based on the ratio of content words to function words in a sentence. The content word density feature uses a pre-populated list of function words (a stopword list) to calculate the ratio of content to function words within each sentence, keeping sentences that meet or exceed the ratio of 50%. This cut-off value implies that the extracted sentences contain relatively more content words than function words.

4.4 Sentence Scoring and Selection

We compute the final score of each sentence by adding the scores obtained for the MEAD default configuration features (centroid and position) to the scores for the COGENT features (educational standards, gazetteer, and hypertext). After the sentences have been sorted according to their cumulative scores, we keep sentences that pass the cut-off constraints, including the MEAD length feature equal or greater than 9 and COGENT content word density equal or greater than 50%. We use the MEAD cosine re-ranker to eliminate redundant sentences based on a cutoff similarity value of 0.7. Since human experts used only 5% of the total word count in the resources, we have configured MEAD to use a 5% word compression rate.

5 Evaluation

We have evaluated COGENT by processing the 20 digital library resources used in the empirical study and comparing its output against the concepts identified by the experts.

5.1 Quality

To assess the quality of the generated summaries, we have examined three configurations: *Random*, *Default*, and *COGENT*. The *Random* configuration extracts a random collection of sentences from the input texts. The *Default* configuration uses the MEAD default centroid, position and length (cut-off value of 9) sentence scoring features. Finally, the *COGENT* configuration includes the MEAD default features and the COGENT features. The Default and COGENT configurations use the MEAD cosine function with a threshold of 0.7 to eliminate redundant sentences. All three configurations use a word compression factor of 5% resulting in summaries of very similar length.

For this evaluation, we leverage ROUGE (Lin and Hovy, 2003) to address the relative quality of the generated summaries based on common n-gram counts and longest common subsequence (LCS). We report on ROUGE-1 (unigrams), ROUGE-2 (bigrams), ROUGE W-1.2 (weighted LCS), and ROUGE-S* (skip bigrams) as they appear to correlate well with human judgments for longer multi-document summaries, particularly ROUGE-1 (Lin, 2004). Table 2 shows the results of this ROUGE-based evaluation including recall (R), precision (P), and balanced f-measure (F).

		Random	Default	COGENT
R-1	R	0.4855	0.4976	0.6073
	P	0.5026	0.5688	0.6034
	F	0.4939	0.5308	0.6054
R-2	R	0.0972	0.1321	0.1907
	P	0.1006	0.1510	0.1895
	F	0.0989	0.1409	0.1901
R-W-1.2	R	0.0929	0.0951	0.1185
	P	0.1533	0.1733	0.1877
	F	0.1157	0.1228	0.1453
R-S*	R	0.2481	0.2620	0.3820
	P	0.2657	0.3424	0.3772
	F	0.2566	0.2969	0.3796

Table 2. Quality evaluation results (5% word compression)

COGENT consistently outperforms the Random and Default baselines based on all four reported ROUGE measures. Given that much of the original research efforts on MEAD have centered on news articles, this result is not surprising. Pedagogical content, such as the educational digital library resources used in our work, differs in rhetorical intent, structure and terminology from the news articles leveraged by the MEAD researchers. However, the COGENT features described here are complementary to the default MEAD configuration. COGENT can best be characterized as a hybrid MDS, integrating bottom-up (centroid, position, length, hypertext, and content word density) and top-down (educational standards and gazetteer) sentence scoring features. This hybrid approach reflects our findings from observing expert behaviors for identifying concepts from educational digital library resources. We believe the overall improvement in quality scores may be due to the COGENT features targeting different dimensions of what constitutes a pedagogically effective summary than the default MEAD features.

To characterize the COGENT summary contents, one of our research team members manually generated a summary corresponding to the best case for an extractive summarizer. This *Best Case* summary comprises the sentences from the digital library resources that align to the concepts selected by the human experts in our empirical study. Since the experts created concepts of varying granularity, this alignment produces the list of sentences that the experts would have produced if they had only selected single sentences to create concepts for their domain knowledge map. This summary comprises 621 sentences consisting of 13,116 words, or about a 9% word compression.

For this aspect of the evaluation, we have used ROUGE-L, an LCS metric computed using ROUGE. The ROUGE-L computation examines the union LCS between each reference sentence and all the sentences in the candidate summary. We believe this metric may be well-suited to reflect the degree of linguistic surface structure similarity between summaries. We postulate that ROUGE-L may be able to account for the explicitly copy-pasted concepts and to detect the more subtle similarities with paraphrased concepts in the expert-generated domain knowledge map. We have also used the content-based evaluation capabilities of MEAD to report on a cosine measure to capture similarity between the candidate summaries and the reference. Table 3 shows the results of this aspect of the evaluation including recall (R), precision (P), and balanced f-measure (F).

		Random (5%)	Default (5%)	COGENT (5%)	Best Case (9%)
R-L	R	0.4814	0.4919	0.6021	0.9669
	P	0.4982	0.5623	0.5982	0.6256
	F	0.4897	0.5248	0.6001	0.7597
Cosine		0.5382	0.6748	0.8325	0.9323

Table 3. Content-based evaluation results (word compression in parentheses)

COGENT consistently outperforms the Random and Default baselines on both the ROUGE-L and cosine measures. Given the cosine value of 0.8325, it appears COGENT extracts sentences containing similar terms in very similar frequency distribution as the experts.

The ROUGE-L scores also consistently indicate that the COGENT summary may be closer to the reference summary in relative word ordering than either the Random or Default configurations. However, the scores for the Best Case summary reveal two interesting points. First, the ROUGE-L recall score for COGENT (R=0. 6021) is lower than that obtained by the Best Case summary (R=0.9669), meaning our summarizer appears to be extracting different sentences than those selected by the experts. Given the high cosine similarity with the reference summary (0.8325), we hypothesize that COGENT may be selecting sentences that cover very similar concepts to those selected by the experts only expressed differently. Second, we would have expected the ROUGE-L precision score for the Best Case configuration to be closer to 1.0. Instead, the Best Case precision score is 0.6256, only a minor improvement over COGENT (P=0.5982). Since the sentences in the Best Case summary come directly from the digital library resources, we hypothesize that experts may have used extensive linguistic transformations for paraphrased concepts, resulting in structures that ROUGE-L could not identify as similar.

Given the difference in word compression for the Best Case summary, we have performed an incremental analysis using the ROUGE-L measure shown in Figure 2.

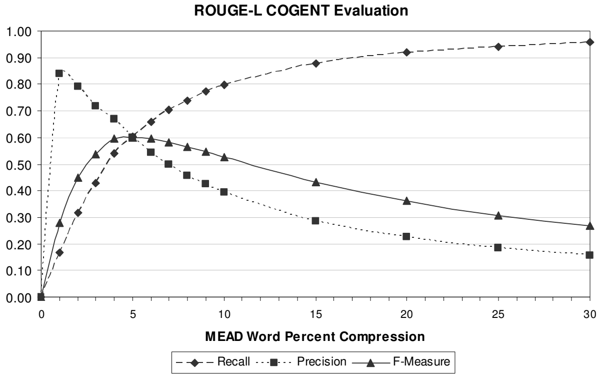


Figure 2. COGENT ROUGE-L results at different word compression rates

This graph shows improved COGENT performance in ROUGE-L recall as the length of the summary increases, while both precision and f-measure degrade. COGENT can match the recall scores of the Best Case summary (R=0.9669) by making the generated summary longer (30% word compression rate or 32,619 words), but the precision would suffer a sizeable decay (P=0.1558). For educational applications, more comprehensive concept inventories (longer summaries) may be better suited for computational purposes, such as pedagogical reasoning about student understanding, while more succinct inventories (shorter summaries) may be more appropriate for display to the student.

5.2 Pedagogical Utility

We have evaluated COGENT's pedagogical utility in the context of computationally identifying student scientific misconceptions. We have developed algorithms that reliably detect incorrect statements in student essays by comparing an expert-created domain knowledge map to an expert-created knowledge map of an essay. These algorithms use textual entailment techniques based on a shallow linguistic analysis of knowledge map concepts to identify sentences that contradict concepts in the domain knowledge map. Initial evaluation results indicate that these algorithms identify incorrect statements nearly as adeptly as human experts.

	Manual Expert Agreement	Expert Knowledge Maps	COGENT Concept Inventory
Recall	0.69	0.87	0.93
Precision	0.69	0.57	0.57
F-Measure	0.69	0.68	0.69

Table 4. Incorrect statement identification evaluation results

As shown in Table 4, the algorithms detect 87% of all incorrect statements identified by experts and 57% of the reported incorrect statements agree with human judgments on the same task. By comparison, experts show 69% overlap on average along both dimensions. Introducing the COGENT concept inventory in place of the expert-created domain knowledge map improves recall performance, as the algorithms return 93% of all incorrect statements reported by the experts, while preserving 57% precision. These results indicate that the generated summary covers the necessary pedagogical concepts to computationally identify student scientific misconceptions.

Informal sampling of the sentences selected by COGENT shows the following three important science concepts receiving the highest scores:

1. Earthquakes are the result of forces deep within the Earth's interior that continuously affect the surface of the Earth.
2. Scientists believed that the movement of the Earth's plates bends and squeezes the rocks at the edges of the plates.
3. In particular, four major scientific developments spurred the formulation of the plate-

tectonics theory: (1) demonstration of the ruggedness and youth of the ocean floor; (2) confirmation of repeated reversals of the Earth magnetic field in the geologic past; (3) emergence of the seafloor-spreading hypothesis and associated recycling of oceanic crust; and (4) precise documentation that the world's earthquake and volcanic activity is concentrated along oceanic trenches and submarine mountain ranges.

For a more rigorous analysis of the pedagogical utility of the COGENT concepts, we asked an instructional expert with domain expertise in geology to evaluate the 326 sentences returned by COGENT. The expert used a 5-point Likert scale to judge whether each concept would be pedagogically useful in the context of a concept inventory on earthquakes and plate tectonics knowledge for high school age learners. The expert agreed or strongly agreed that 60% of the sentences would be pedagogically useful, with 30% of the sentences being potentially useful and only 10% of the sentences being judged as not useful. These results indicate that COGENT appears to perform quite well at identifying sentences that contain information relevant for learning about the domain.

We have also completed an ablation study to identify the relative contribution of the COGENT features to the quality of the summary. We have focused on the cosine metric to capture the overall similarity between the COGENT concept inventory and the concepts from the expert-created knowledge map.

Features	Cosine
All Features	0.8325
(Gazetteer)	0.5545
(Hypertext)	0.5575
(Educational Standards)	0.8083
(Content Word Density)	0.8271

Table 5. Feature ablation evaluation results for COGENT

Table 5 shows the cosine similarity between the concept inventory generated after taking the feature shown in parentheses out of the summarizer. The results are ordered from low-to-high such that the feature contributing the most to the all-features cosine score appears at the top of the table. Removing either the gazetteer or the hypertext feature causes the largest drops in similarity indicating the importance of the use of examples and the relevance of document structure for the quality of the COGENT-generated summary. Meanwhile both the educational standards and content word density appear to provide modest but useful improvements to the quality of the COGENT summary.

Given that our algorithms have only been evaluated on the topic of earthquakes and plate tectonics for high school age learners, COGENT may be limited in its ability to transcend domains due to its reliance on two domain-aware sentence scoring features: educational standards and gazetteer. However, the educational standards feature may be applicable across other science topics because the *AAAS Benchmarks* and NSES provide very thorough and detailed coverage of a wide range of topics for the Science, Technology, Engineering, and Math disciplines for grades K-12. Only the gazetteer feature would need to be replaced, especially given its significant contribution to the quality of the generated summary as indicated by the results of the ablation study. We believe these results highlight the need to generalize our approach, perhaps using a classifier for identifying examples in educational texts without resorting to overly domain-specific language resources, such as the ADL Gazetteer. Overall, the evaluation results indicate that our approach holds promise for effectively identifying concepts for inclusion in the construction of a pedagogically useful domain knowledge map from educational science content.

6 Conclusions and Future Work

In this paper, we have presented a multi-document summarization system, COGENT, that integrates bottom-up and top-down sentence scoring features to identify pedagogically relevant concepts from educational digital library resources. Our results indicate that COGENT generates concept inventories that resemble those identified by experts and outperforms existing multi-document summarization systems. We have also used the COGENT concept inventory as input to our misconception identification algorithms and the evaluation results indicate the algorithms perform as well as when using an expert-created domain knowledge map. In the context of generating domain knowledge maps, our next step is to explore how machine learning techniques may be employed to connect concepts with links.

Automating the process of creating inventories of important pedagogical concepts represents an important step towards creating scalable intelligent learning and tutoring systems. We hope our progress in this direction may contribute to increase the interest within the computational linguistics research community in novel educational technology research.

Acknowledgments

This research is funded in part by the National Science Foundation under NSF IIS/ALT Award 0537194. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

Ahmad, F., de la Chica, S., Butcher, K., Sumner, T. and Martin, J.H. (2007, June 17-23). Towards automatic conceptual personalization tools. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, (Vancouver, British Columbia, Canada, 2007), pages 452 - 461.

Butcher, K.R., Bhushan, S. and Sumner, T. (2006). Multimedia displays for conceptual discovery: information seeking with strand maps. ACM Multimedia Systems, 11 (3), pages 236-248.

Byrt, T., Bishop, J. and Carlin, J.B. (1993). Bias, prevalence, and kappa. Journal of Clinical Epidemiology, 46 (5), pages 423-429.

Harabagiu, S. and Lacatusu, F. (2005, August 15-19). Topic themes for multi-document summarization. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (Salvador, Brazil, 2005), pages 202-209.

Hardy, H., Shimizu, N., Strzalkowski, T., Ting, L., Wise, G.B. and Zhang, X. (2002). Summarizing large document sets using concept-based clustering. In Proceedings of the Human Language Technology Conference 2002, (San Diego, California, United States, 2002), pages 222-227.

Hill, L.L. (2000, September 18-20). Core elements of digital gazetteers: placenames, categories, and footprints. In Proceedings of the 4th European Conference on Digital Libraries, (Lisbon, Portugal, 2000), pages 280-290.

Holley, C.D. and Dansereau, D.F. (1984). Spatial learning strategies: Techniques, applications, and related issues. Academic Press, Orlando, Florida.

Lin, C.Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out, (Barcelona, Spain, 2004).

Lin, C.Y. and Hovy, E. (2003, May-June). Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL, (Edmonton, Canada, 2003), pages 71-78.

Lin, D. and Pantel, P. (2002, August 24-September 1). Concept discovery from text. In Proceedings of the 19th International Conference on Computational Linguistics, (Taipei, Taiwan, 2002), pages 1-7.

Mani, I. (2001). Automatic Summarization. Mitkov, R. (Ed.) John Benjamins B.V., Amsterdam, The Netherlands.

National Research Council. (1996). National Science Education Standards. National Academy Press, Washington, DC.

Navigli, R. and Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated websites. Computational Linguistics, 30 (2), pages 151-179.

Novak, J.D. and Gowin, D.B. (1984). Learning how to learn. Cambridge University Press, New York, New York.

O'Donnell, A.M., Dansereau, D.F. and Hall, R.H. (2002). Knowledge maps as scaffolds for cognitive processing. Educational Psychology Review, 14 (1), pages 71-86.

Project 2061. (1993). Benchmarks for science literacy. Oxford University Press, New York, New York, United States.

Radev, D.R., Jing, H. and Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In Proceedings of the ANLP/NAACL 2000 Workshop on Summarization, (2000), pages 21-30.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24 (5), pages 513-523.

Wan, X. and Yang, J. (2006, June 5th-7th). Improved affinity graph based multi-document summarization. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, (New York City, New York, 2006), pages 181-184.

Zouaq, A., Nkambou, R. and Frasson, C. (2007, July 9-13). Learning a domain ontology in the Knowledge Puzzle project. In Proceedings of the Fifth International Workshop on Ontologies and Semantic Web for E-Learning, (Marina del Rey, California, 2007).