DesignWebs: A Tool for Automatic Construction of Interactive Conceptual Maps from Document Collections

Sharad V. Oberoi, Dong Nguyen, Gahgene Gweon,
Susan Finger, and Carolyn Penstein Rosé

Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA, 15213
{svo,dongn,gkg,sfinger,cp3a}@andrew.cmu.edu

Abstract. Prior work supports the pedagogical value of conceptual maps for offering students an overview of a topic as well as the connections between subtopics. In this poster we describe a system that uses automated topic modeling technology to map the topics and sub-topics in a collection of documents. An interactive graphical representation allows users to explore this topic analysis, using it as an interface for browsing a collection of documents. We present a small user study evaluating the usability of the interactive map.

Keywords: conceptual maps, graphical representations, language processing.

1 Introduction

This poster presents a framework that supports knowledge management for project teams. An important part of many group projects is becoming aware of the state-of-the art in areas relevant to the design or development goal. As supporters of the learning process, instructors can raise student awareness of relevant knowledge and support student integration of that knowledge. However, it may still occur with their limited experience that students miss important connections between concepts. We describe a working system that constructs conceptual maps automatically from document collections, such as from the articles found through Google Scholar. These conceptual maps are referred to as DesignWebs [1]. DesignWebs are also a navigation aid since the nodes in the map link directly to the source documents.

2 Creating DesignWebs

Providing concept maps to students has been suggested as a metacognitive tool to enhance their learning in the sciences [2,3].Topic modeling approaches identify high-level topics present in a document collection. One such method is Latent Dirichlet Allocation (LDA)[4], which is a generative process that models each document as a mixture of topics, and models each topic as a multinomial distribution over words. To construct a DesignWeb, the instructor needs to gather a collection of relevant documents and start the automatic preprocessing script. The result is a DesignWeb not

V. Aleven, J. Kay, and J. Mostow (Eds.): ITS 2010, Part II, LNCS 6095, pp. 387–389, 2010.

only showing the main topics within a given set of documents, but also allowing users to zoom-in to view sub-topics and the terms relevant to these topics (See Figure 1). Users can also browse the relevant documents.

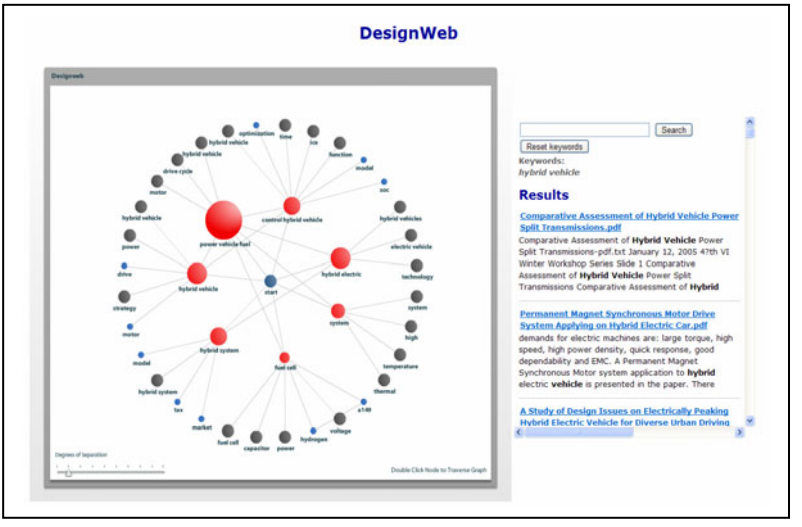

Fig. 1. Screenshot of the DesignWeb interface

The main steps of our technical approach are:

1. We first apply LDA to model the topics in a document collection. After the extraction of the topic model, a hierarchy for every topic is induced using hierarchical clustering.
2. Term lists associated with topics were compressed by collapsing terms that had similar topic distributions and could be reduced to the same stem after removing any morphological endings. In order to include common phrases in the vocabulary list, we extract the documents most strongly associated to each topic and then extract the collocations from this set. Clusters are labeled by counting unigrams and bigrams in the set of most relevant documents of the cluster and choosing the label the one that has the highest count.
3. Connections between topics are computing by first representing each topic as a vector of probabilities that represent the topic's associated word distribution. Cosine similarity is then computed for each pair of topics, and if it is higher than a threshold (0.2), the topics are considered to be linked.
4. Each node is associated with a topic in the topic model.

Clicking on a node in the Design Web issues a query that retrieves the documents most strongly associated with the topic. This query is treated as though it contains all the terms most strongly associated with the topic so that when the documents are retrieved, and snippets are displayed for each, a snippet will be selected that contains a high concentration of those terms. Additional terms can be added to this query in

order to influence the displayed snippets. We use the Lingpipe framework[1] for extracting the LDA model, collocations and applying the clustering. Lucene[2] is used for the document retrieval functionality.

3 Informal Evaluation and Current Work

We conducted a small user study to evaluate how well students are able to use Design Webs to explore a document collection. The scenario for our user study is a class of engineering graduate students about to embark on a project involving an analysis of hybrid cars. To simulate a typical a literature review task at the start of such a class, we downloaded 250 articles from Google Scholar and automatically generated a DesignWeb from these documents. The students were given a demonstration of the system using a DesignWeb created from a different corpus and allowed to acquaint themselves with it. Then, they were asked to identify alternative battery types and alternative energy sources for hybrid cars from the DesignWeb, along with the pros and cons of each. The students had to cite the sources and not use any prior knowledge. The students' answers for the alternative battery types and energy sources varied between 3 and 7 choices each, with an average of 5 choices per student. The number of sources cited varied between 3 and 10 (average: 5.4), while the total number of advantages and disadvantages cited was between 5 and 10 (average: 7.4).

DesignWebs provide a robust and automatic method to organize, navigate and synthesize the documents referenced by students during a project. These are expected to support learning tasks by providing a bird's eye-view that is otherwise not possible due to information scattered in research literature that is needed for the task.

This work was supported by NSF Grants EEC-0935127 and EEC-064848.

References

1. Oberoi, S.V., Finger, S.: DesignWebs: An Interactive Organizational Memory Assimilation and Navigation Tool. In: 17th International Conference on Engineering Design, Stanford, CA, August 24-27 (2009)
2. Horton, P.B., McConney, A.A., Gallo, M., Woods, A.L., Senn, G.J., Hamelin, D.: An Investigation of the Effectiveness of Concept Mapping as an Instructional Tool. Science Education 44, 95–111 (1993)
3. Lawless, C., Smee, P., O'Shea, T.: Using Concept Mapping and Concept Mapping in Business and Public Administration, and in Education: An Overview. Educational Research 40(2), 219–235 (1998)
4. Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical Topic Models and the Nested Chinese Restaurant Process. In: Advances in Neural Information Processing Systems, p. 2003 (2004)

[1] Lingpipe: http://alias-i.com/lingpipe/
[2] Lucene: http://lucene.apache.org/