Discovering Concept Maps from Textual Sources

R.P. Jagadeesh Chandra Bose Om Deshmukh B. Ravindra

Xerox Research Center India
Etamin Block 3, 4th Floor, Wing-A, Prestige Tech Park II, Bangalore, India 560103.

{jagadeesh.prabhakara, om.deshmukh}@xerox.com

ABSTRACT

Concept maps and knowledge maps, often used as learning materials, enable users to recognize important concepts and the relationships between them. For example, concept maps can be used to provide adaptive learning guidance for learners such as path systems for curriculum sequencing to improve the effectiveness of learning process. Generation of concept maps typically involve domain experts, which makes it costly. In this paper, we propose a framework for discovering concepts and their relationships (such as prerequisites and relatedness) by analyzing content from textual sources such as a textbook. We present a prototype implementation of the framework and show that meaningful relationships can be uncovered.

1. INTRODUCTION

In any given learning setting, a hierarchy of concepts (set by experts) is provided and the learner is expected to follow through these concepts in the specified order, e.g., Table of Contents (ToC), which indicates that concepts appearing in earlier chapters *are* (sometimes *'may be'*) pre-requisites for the concepts discussed in the later chapters. Similarly, end-of-the-book index indicates prominent occurrences of the main concepts (and some relationships between them) discussed in the book. In both the cases, the relationship is static, is designed by the experts and is restricted to the pre-populated list of concepts. As we move towards personalized learning, such a knowledge-driven static elicitation is inadequate. e.g., if the immediate goal of the learner is to understand concepts in chapter L, s/he may only have to go through a select 'n' sections of some chapters till L. Consider another example, if a learner has to know which concepts co-occur or which concepts predominantly occur before a particular concept C and are relevant to the concept C. This information is not easily available either from the ToC or from the "end-of-the-book index".

Concept map is a knowledge visualization tool that represents concepts and relationships between them as a graph. Nodes in the graph correspond to concepts and edges depict the relationship between concepts. In recent years, concept maps are widely used for facilitating meaningful learning, capturing and archiving expert knowledge, and organizing and navigating large volumes of information. In adaptive learning, concept maps can be used to give learning guidance by demonstrating how the learning status of a concept can possibly be influenced by learning status of other concepts [3]. Construction of concept maps is a complex task and typically requires manual effort of domain experts, which is costly and time consuming.

In this paper, we propose a framework for automatic generation of concept maps from textual sources such as a textbook and course webpages. We discover concepts by exploiting the structural information such as table of contents and font information and establish how closely two concepts are related to each other where the relation is defined on how strongly one concept is being referred to/discussed in another. The proposed approach is implemented and applied on several subjects. Our initial results indicate that we are able to discover meaningful relationships.

The remainder of this paper is organized as follows. Related work is presented in Section 2. We discuss our approach of discovering concept maps in Section 3. Section 4 presents some experimental results. Section 5 concludes with some directions for future work.

2. RELATED WORK

Concept map mining refers to the automatic or semi-automatic creation of concept maps from documents [4]. Concept map mining can be broadly divided into two stages: (i) concept identification and (ii) concept relationships association. Concept identification is typically done using dictionaries or statistical means (e.g., frequent words). Relation between concepts is typically defined over word-cooccurrences. In our work, we do not use any dictionary of terms. Instead, we rely on structural information such as bookmarks, table of contents, and font information manifested in data sources to discover concepts. Furthemore, when discovering relationships, we not only look at co-occurence of concepts within a sentence but scope it to larger segments such as a section and chapter.

3. GENERATION OF CONCEPT MAPS

Concept maps should provide support for modular nature of the subject matter and the interconnections between knowledge modules (concepts). Formally, a concept map can be defined as a tuple $< C, R, L >$ where $C = \{c_1, c_2, \ldots, c_n\}$ is a set of concepts; $L = \{l_1, l_2, \ldots, l_k\}$ is a set of labels. $R = \{r_1, r_2, \ldots, r_m\} \subseteq C \times C \times L$ is a set of relationships among concepts. Each relation $r_j = (c_p, c_q, l_s) \in R, p \neq q, 1 \leq p, q \leq n, 1 \leq j \leq m, 1 \leq s \leq k$ defines a relation-

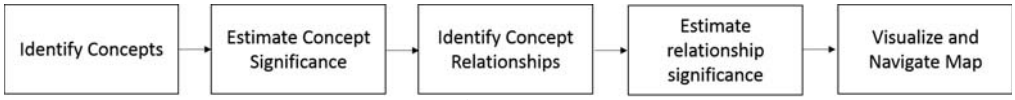

Figure 1: Approach Overview

ship between concept c_p and c_q which is labeled l_s. Optionally each relation r_j can also be associated with a weight $w_j \in \Re^+$. Figure 1 presents an overview of our approach and is comprised of five steps:

1. Identify Concepts: We exploit structural and font information such as bookmarks, table of contents, and index (glossary) in e-textbooks, and headers and font information in html pages for this step. Text processing such as tokenizing, stemming, and stop word removal are then applied. Concepts are identified as either individual words or n-words ($n > 1$)

2. Estimate Concept Significance: We estimate the significance of concepts automatically using different criteria: (i) frequency of occurrence (frequent concepts are more significant than infrequent ones) (ii) importance of a concept w.r.t the examinations/evaluations and (iii) font related information (larger font concepts are more significant than smaller fonts). The three criteria mentioned above can be grouped together using weights.

3. Identify Concept Relationships: Several types of relationships can be defined among concepts, e.g., superclass-subclass (one concept is *more general* than another), prerequisite relation (a concept A is said to be a pre-requisite for concept B), etc. The table of contents in a document directly gives a (partial) hierarchical structure among concepts. Apart from the hierarchical relationship, concepts can also be horizontally related e.g., *relevant to* and *mentioned by* as discussed in [1]. We consider the *mentioned by* relation, which is used to express the fact that two concepts are related of the type A *refers-to* B, A *discusses* B, A *mentions* B. Note that *mentioned by* is an *asymmetric* and *not necessarily transitive* relation.

4. Estimate Relationship Significance: Relationship significance is estimated using *term co-occurrence* as a basis. For each concept, in the pages where it manifests, we also estimate which other concepts manifest in those pages and how often do they manifest. The degree of relatedness is obtained by the frequency at which the concept is used, e.g., if concept c_j manifests f_j times when describing concept c_i and if f_i is the frequency of occurrence of concept c_i, then the weight of the edge between c_j and c_i can be defined as f_j/f_i. We also consider normalized weights.

5. Visualize and Navigate Map: The concepts and their relationships can be visualized as a graph $G = (V, E)$ where V, the set of vertices, correspond to the concepts and E, the set of edges, correspond to the relationship between concepts. Nodes and edges can be annotated to provide rich information and enable the navigation of these maps e.g., size of the node can be used to depict the significance of a concept, color of the node can be used to indicate its importance w.r.t student examinations/evaluation, thickness of the node can be used to depict the relative knowledge of the student on the concept. Similarly, edges can be annotated to reveal different kinds of information e.g., thickness of an edge can be used to signify the relatedness between two concepts.

4. EXPERIMENTS AND DISCUSSION

We have implemented the proposed framework in Java and Python and tested it on several examples. Visualization of concept maps is implemented using d3js. In this section, we present the results of one such experiment of generating concept maps using the pdf textbook on databases [2]. Figure 2 depicts a subgraph corresponding to the concepts related to relational algebra. We showed the uncovered concept maps

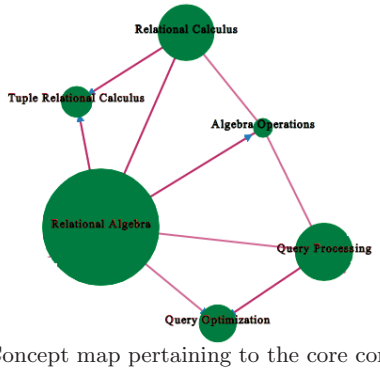

Figure 2: Concept map pertaining to the core concept relational algebra

to a few experts in databases and they mostly agree to the discovered relations. We have applied our approach to several subjects (e.g., operating systems, computer networks etc.) and found that in each of those, we are able to uncover meaningful and important relations. We realize that there is a need for an objective evaluation method to automatically assess the goodness of discovered concept maps, e.g., using gold standard.

5. CONCLUSIONS

Generation of concept maps is an important means of supporting deep understanding of a subject matter. In this paper, we presented an approach for identifying concepts and establishing how closely two concepts are related to each other. We believe that these concept maps enable users to quickly get knowledge about the centrality or importance of each concept and its significance in understanding other concepts. As future work, we would like to further enrich the discovered concept maps with additional information based on the user of the application. For example, upon clicking on a node, teachers/faculty can be provided with information such as the average/distribution score of students on this concept in various tests conducted; students can be provided with links to lecture material, questions/solutions asked in previous exams, etc.

6. REFERENCES

[1] Darina Dicheva and Christo Dichev. Authoring educational topic maps: can we make it easier? In *ICALT*, pages 216–218, 2005.

[2] R. Elmasri and S.B. Navathe. *Fundamentals of database systems*. Pearson Education India, 6 edition, 2010.

[3] Shian-Shyong Tseng, Pei-Chi Sue, Jun-Ming Su, Jui-Feng Weng, and Wen-Nung Tsai. A new approach for constructing the concept map. *Computers & Education*, 49(3):691–707, 2007.

[4] Jorge J. Villalon and Rafael A. Calvo. Concept map mining: A definition and a framework for its evaluation. WI-IAT '08, pages 357–360, 2008.