Expert Systems with Applications 39 (2012) 12709–12718

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

The automatic creation of concept maps from documents written using morphologically rich languages

Krunoslav Zubrinic [a,*], Damir Kalpic [b], Mario Milicevic [a]

[a] Department of Electrical Engineering and Computing, University of Dubrovnik, Cira Carica 4, HR 20000 Dubrovnik, Croatia
[b] Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, Zagreb, Croatia

ARTICLE INFO

Keywords:
Concept map
Concept map mining
Text analysis
Text summarization

ABSTRACT

Concept map is a graphical tool for representing knowledge. They have been used in many different areas, including education, knowledge management, business and intelligence. Constructing of concept maps manually can be a complex task; an unskilled person may encounter difficulties in determining and positioning concepts relevant to the problem area. An application that recommends concept candidates and their position in a concept map can significantly help the user in that situation. This paper gives an overview of different approaches to automatic and semi-automatic creation of concept maps from textual and non-textual sources. The concept map mining process is defined, and one method suitable for the creation of concept maps from unstructured textual sources in highly inflected languages such as the Croatian language is described in detail. Proposed method uses statistical and data mining techniques enriched with linguistic tools. With minor adjustments, that method can also be used for concept map mining from textual sources in other morphologically rich languages.

1. Introduction

Concept map (CM) is a graphical tool that has been successfully used for organizing and representing knowledge. It includes concepts, usually signified by nouns or noun phrases, and relationships between them, indicated by a line linking two concepts. Labelling a line with a verb or a verb phrase creates a concept-label-concept chain that can be read as a sentence. This chain is called a proposition (Novak & Cañas, 2008).

There has been a remarkable growth in the use of CMs throughout the world over the past decade. The most prevalent applications of concept mapping are facilitating meaningful learning, and capturing and archiving expert knowledge in a form that would be easy to use by others. Furthermore, CMs have been known to be an effective tool to organize and navigate through large volumes of information.

In personal learning, a CM can be used as the tool that represents a learning plan, which consists of a set of goals that a person hopes to achieve within a specific period. For most learners it is difficult to begin with a “blank sheet” and start to build a map for a chosen topic of interest. A skeleton map provided by an expert can make it easier for the learner to start that process. In personal learning, it is difficult to find experts for specific learning fields. Therefore, an information system that behaves like an expert and provides the skeleton of a CM can be very helpful in such situations.

The automatic or semi-automatic creation of CMs from documents is called concept map mining (CMM) (Villalon & Calvo, 2008). In a semi-automatic process, the system finds and suggests elements of a map, and a person has to complete the map manually, using the provided information. In the automatic construction process, the user’s assistance is not required, and the process creates the map automatically from available resources.

This paper introduces research that addresses the automatic creation of a CM from unstructured text in the morphologically rich Croatian language. It describes the first stage of that research, and presents its threefold direction: (a) to gain a better understanding of the problem area, (b) to collect information and materials relevant to the research problem, and (c) to identify convenient technologies and procedures that can be used in later phases of the research.

This paper is structured as follows. In the second chapter, the CM and CMM-related terms are defined. Literature review of different approaches and previous work related to CMM is given in the third chapter. A procedure for the CMM of unstructured textual documents in the Croatian language is proposed and described in the fourth chapter. A short discussion of proposed method is given in the fifth chapter. The sixth chapter provides a brief summary of the paper, and presents a plan for future research activities.

* Corresponding author. Tel.: +385 20 445 742; fax: +385 20 445 770.
E-mail addresses: krunoslav.zubrinic@unidu.hr (K. Zubrinic), damir.kalpic@fer.hr (D. Kalpic), mario.milicevic@unidu.hr (M. Milicevic).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.04.065

2. Concept map mining

This chapter explains and formally defines elements of the CM and the main terms related to a CMM process.

2.1. Concept map

Semantic network is a formal structure for representing knowledge as a pattern of interconnected nodes and arcs. Its notation is powerful enough to represent the semantics of natural languages, and can be automatically processed by computer programs (Kramer & Mylopoulos, 1987). CM is a special type of propositional semantic network that is flexible and oriented to humans. It is designed in the form of a directed graph where nodes represent concepts and arcs represent relationships among them (McNeese et al., 1990).

The educational technique of concept mapping was first attributed to education theorist Novak in 1970s, when his group of researchers described the human learning process as a lifelong process of assimilating new concepts and relations into a personal conceptual framework (Novak & Cañas, 2008). Novak adopted the semantic network model and created the CM as a tool for the graphical representation of a learner's conceptual understanding of information in a specific area. His initial idea was that a CM should be drawn free hand by a learner, but only after an initial articulation of major ideas and their classification in hierarchical manner. We now understand, however, that the topology of a CM can take a variety of forms, ranging from hierarchical, to non-hierarchical and data-driven forms.

Formally, a hierarchical CM can be defined (Villalon & Calvo, 2008) as a set

$$CM = \{C, R, T\}$$

where

- $C = \{c_0, c_1, \ldots, c_{n-1}\}$ is a set of concepts. Each concept $c_i \epsilon C; 0 \leqslant i < n$ is a word or phrase, and is unique in C.
- $R = \{r_0, r_1, \ldots, r_{m-1}\}$ is a set of relationships among concepts. Each relationship $r_j \epsilon R = (c_p, c_q, l_j); p \neq q; 0 \leqslant p < n; 0 \leqslant q < n; 0 \leqslant j < m$, connects two concepts, c_p, $c_q \in C$. Label l_j is a term that labels relationship r_j and represents a conceptual relationship between coupled concepts.
- $T = \{t_0, t_1, \ldots, t_{s-1}\}$; $t_{k-1} < t_k < t_{k+1}$; $0 < k < s - 1$ is a sorted set of hierarchical levels in a CM. Each element $t_k \epsilon T = \{c_0, c_1, \ldots, c_{r-1}\}; 0 \leqslant r < n$ corresponds to a set of concepts that share the same level of generalization in a CM.

2.2. Definition of concept map mining

CMM is a process of extracting information from one or more documents for the automatic creation of a CM. The created map should be a generic summary of a source text (Villalon & Calvo, 2008).

From the CMM point of view, a document can be formalized as a set

$$D = \{C_d, R_d\}$$

where

- $C_d = \{c_{d0}, c_{d1}, \ldots, c_{dn-1}\}$ is a set of all concepts, and
- $R_d = \{r_{d0}, r_{d1}, \ldots, r_{dm-1}\}$ is a set of all relationships that can be extracted from the document.

Three general phases of the CMM process (Villalon & Calvo, 2008) are depicted in Fig. 1.

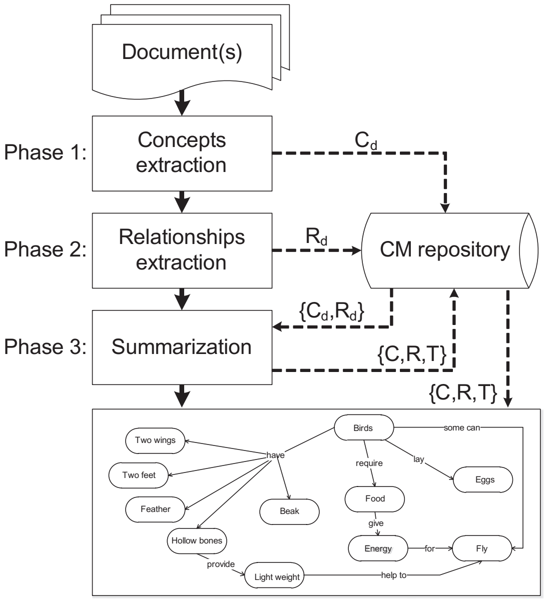

Fig. 1. General CMM process.

The first phase is the identification and extraction of all members of the set C_d. These members are concepts, represented by subjects and objects in the text—usually nouns or noun phrases. For correct and complete understanding of concepts in a map, it is important to consider their connection with synonyms and homonyms. In the natural language processing (NLP) field, that information is usually stored in a repository called a lexicon, where lexical terms are connected with their syntactic or semantic properties. In CMM the same approach can be used, that is, the connection of a concept with its synonyms can be recorder in a simple lexicon. That lexicon can be formalized as a set $Lex = \{lex_0, lex_1, \ldots, lex_{n-1}\}$ where each element is a pair $lex_i = (c_i, S_i)$ that contains a concept and a set of concept's synonyms $S_i = \{s_0, s_1, \ldots, s_{m-1}\}$.

Noun can be used as a subject or object in a sentence. When syntactic or semantic dependency between subject and object in a sentence is known, it is possible to extract link that exists between them. Extracting links is a goal of the second phase in this process. A chosen relationship becomes member of the set R_d. Each member of that set $r_{dj} \epsilon R_d = (c_{dp}, c_{dq}, l_{dj}); dp \neq dq; 0 \leqslant dp < dn; 0 \leqslant dq < dn; 0 \leqslant dj < dm$ connects two concepts, c_{dp}, $c_{dq} \in C_d$ in a document. Third member of the set r_{dj} is text label l_{dj}, which labels a connection between concepts c_{dp} and c_{dq}. The final phase of the CMM process is a summarization of the information extracted from document D and the creation of a set $CM = \{C, R, T\}$ that contains concepts, relationships and topological information of the map.

The goal of the CMM process is to produce a CM that is an accurate visual abstract of a source text. That visualization is intended for human analysis, and should not contain too many concepts, preferably between 15 and 25 (Novak & Cañas, 2008). In an educational context, the terminology used in a document is important for users, so the CM should be represented using terms that the author uses in the original text.

The source of the CMM technique can be traced back to the early work of Trochim, who proposed the concept mapping process that combines a group activity with statistical analyses (Trochim, 1989). During a brainstorming session the group of participants creates a set of statements relevant to the domain of interest. Each participant sorts and rates every statement, creating an individual similarity matrix. All personal matrices are then used to form a group proximity array. The most important statements are chosen using a multidimensional scaling (MDS) and hierarchical cluster analysis. This approach, based on weight calculation and statistical and data mining techniques is still commonly used in many contemporary CMM methods.