Automatic creation of a concept map

Krunoslav Žubrinić
University of Dubrovnik, Department of electrical engineering and computing
Ćira Carića 4, Dubrovnik, Croatia
Email: krunoslav.zubrinic@unidu.hr

***Abstract*—Concept map is a graphical technique for representing knowledge, successfully used in different areas, including education, knowledge management, business and intelligence. In this paper, an overview of different approaches to automatic creation of concept maps from textual and non-textual sources is given. Concept map mining process is defined, and one method for creation of concept maps from unstructured textual sources in the Croatian language is described.**

***Index Terms*—concept map, concept map mining, text analysis, text summarization**

I. Introduction

There has been a remarkable growth in the use of concept maps (CMs) across the world over the past decade. The most prevalent application of concept mapping is in facilitating meaningful learning and as a tool for 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 a 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 chosen topic of interest. A skeleton map provided by an expert can help learner in easier starting of a learning process. In personal learning, it is hard to find an expert for specific learning field, and an information system that replaces the expert and provides the skeleton of a CM can be very helpful in that situation.

CM is a graphical tool 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 [1].

Automatic creation of CMs from documents is called concept map mining (CMM) [2]. CMM can be semi-automatic or completely automatic. In the semi-automatic process, the system finds and suggests some elements of a map, and a person manually has to finish the map, using provided information. In automatic construction, the user's assistance is not required, and the process creates the map automatically from available resources.

Introduction into research that addresses the problem of automatic creation of a CM from unstructured text in the Croatian language is given in this paper. It describes the first stage of that research, and its purpose is threefold: a) to gain a better understanding of the research area; b) to collect information and materials relevant to the research problem; and c) to identify potentially feasible technologies that could be used in later phases of the research.

This paper is structured as follows. In the second chapter, the CM and CMM-related terms have been defined. Literature review of different approaches to CMM is given in the third chapter. A procedure for CMM of unstructured textual documents in the Croatian language is proposed and described in the fourth chapter. The fifth chapter is the conclusion where a brief summary of the paper, and plans for future research are given.

II. Concept map mining

In this chapter, main terms related to a CMM process are explained and formally defined.

A. Concept map

Semantic network is a structure for representing knowledge as a pattern of interconnected nodes and arcs. Its notation is efficient for computers to process, and powerful enough to represent the semantics of natural languages [3]. CM is a special type of a 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 [4].

The educational technique of concept mapping has been attributed to education theorist Novak in 1970s, when his group of researchers described the human learning process as a lifelong process of assimilation of new concepts and relations into personal conceptual framework [1]. Novak adopted the semantic network model and created CM as a tool for graphical representation of learner's conceptual understanding of information in a specific area. By initial idea, a CM should be drawn free hand by a learner, after an initial articulation of major ideas and their classification in hierarchical manner. 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 [2] as a set

$$CM = \{C, R, T\} \tag{1}$$

where

$C = \{c_1, c_2, \ldots, c_n\}$ is a set of concepts. Each concept $c_i \in C; 1 \leq i \leq n$ is a word or phrase, and it is unique in C.

$R = \{r_1, r_2, \ldots, r_m\}$ is a set of relationships among concepts. Each relationship $r_i \in R = (c_p, c_q, l_j); p \neq q; 1 \leq p \leq n; 1 \leq q \leq n; 1 \leq j \leq m$, connects two concepts $c_p, c_q \in C$. Label l_j is a term which labels relationship r_j.

$T = \{t_1, t_2, \ldots, t_s\}; t_{k-1} < t_k < t_{k+1}; 1 < k < s$ is a sorted set of hierarchic levels in a concept map. Each element $t_k \in T = \{c_1, c_2, \ldots, c_r\}; 1 \leq r \leq n$ corresponds to a set of concepts that share the same level of generalization in a CM.

B. Definition of concept map mining

CMM is a process of extracting information from one or more documents for automatic creation of a CM. Created map is a generic summary of a source text [2].

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

$$D = \{C_d, R_d\} \tag{2}$$

where

$C_d = \{c_{d1}, c_{d2}, \ldots, c_{dn}\}$ is a set of all concepts and

$R_d = \{r_{d1}, r_{d2}, \ldots, r_{dm}\}$ is a set of all relationships that can be extracted from the document.

Three general phases of the CMM process [2] are depicted in the Fig. 1.

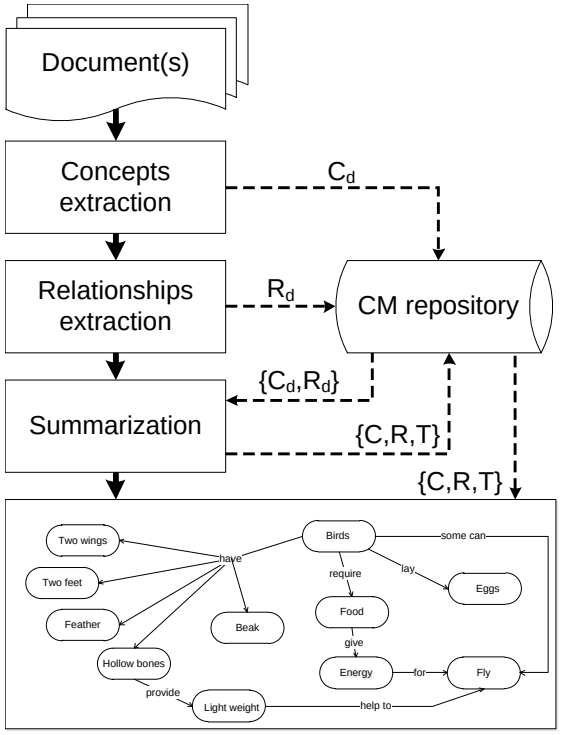

Fig. 1. General CMM process

The first phase is 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. 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. Chosen relationship becomes member of the set R_d. Each member of that set $r_{dj} \in R_d = (c_{dp}, c_{dq}, l_{dj}); dp \neq dq; 1 \leq dp \leq dn; 1 \leq dq \leq dn; 1 \leq dj \leq dm$, connects two concepts, $c_{dp}, c_{dq} \in C_d$ in a document, and label l_{dj} is a term, which labels that relationship. The final phase of the CMM process is a summarization of extracted document's set D and 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 CM that is an accurate visual abstract of a source text. Created map is intended for human analysis, and it should not contain too many concepts, preferably 15–25 [1]. In educational context, the terminology used in a document is important for users, so the CM should be represented using terms that the author used in the original text.

The source of the CMM technique can be traced back to the early work of Trochim, who proposed a concept mapping process that combines a group activity with statistical analyses [5]. The group of participants during brainstorming session creates a set of statements relevant to the domain of interest. Each participant sorts and rates every statement, creating individual similarity matrix. All personal matrices are summed together into 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, statistical and data mining techniques is still commonly used in many contemporary CMM methods.

III. Review of concept map mining approaches

A number of studies have focused on automatic generation of CMs and similar representations from structured and unstructured sources. Some researchers follow strict Novak's definition of a hierarchical CM, while others use knowledge representations that are more-or-less different. A CMM process is mostly supported by methods used in the natural language processing (NLP) field, and these methods are briefly described in the first part of this chapter. In the following parts, a short overview of current CMM studies is given.

A. Methods used in concept map mining

A CMM process can be carried out by NLP methods used in tasks such as information extraction (IE), information retrieval (IR) and automatic summarization. IE is the process of automatic extraction of structured information, such as entities and relations, from unstructured textual sources. IR is area concerned with searching for information in documents and metadata about documents, while the goal of automatic summarization is to distil content from a source, and present the most important content to the user in a condensed form [6]. Summarization result can be extract or abstract. An extract is a summary made of important text segments pulled out from the original document without any changes, while an abstract is a paraphrased summary [7].

Methods traditionally used in these areas are rule-based statistical and machine learning methods. More recently, there has been interest in combining finite-state machines with conditional-probability models, like maximum entropy Markov models and conditional random fields [6]. Most of classical summarization methods are likewise numerical, and based on a weighting model where system weights text elements according to simple word or sentence features, or statistical significance metrics like term frequency-inverse document frequency (TF-IDF). Machine learning methods often provide accurate extracting based on classification, using binary or fuzzy logic. Such method can be used as a main method, or in hybrid systems to supply resources to other processes. Contemporary approaches include hybrid techniques with usage of algorithms in combination with third-party party datasets [7], [8], summarization based on fuzzy logic and swarm intelligence [9].

In NLP field, numerical methods can be enriched with dictionaries of terms or linguistic tools and techniques [6]. Problem with dictionaries arises from the fact that dictionary is usually an external resource. It has to be previously created for

specific domain and it requires further operations for handling new content. A limiting factor for use of linguistic techniques is that appropriate tools and methods are not available for many languages.

B. Mining from unstructured text

Unstructured texts in natural languages are commonly used in a CMM process. Considering the number of documents used for one CM creation, there are two groups of techniques. The first group contains techniques that create one CM from a single document [10]–[19]. Multiple documents are used as a source in the second group of techniques [20]–[29]. Source documents are usually in shorter forms, such as abstracts or full academic papers, student essays, news articles and medical diagnosis reports. Lengthy documents used in CMM researches include theses and dissertations. In adaptive learning environments, few researchers analysed students' learning outcomes using CMs automatically created from students' exam results [30]–[34].

The goal of most studies in this area is to produce a starting CM model, which can speed up the process of CM creation for later refinement by a person, or by another automatic process. Some of created maps are fully completed and contain concepts connected with labelled relationships [11], [12], [14]–[19], [23]–[27]. Other researchers created CMs with connected concepts, but without labelled relationships [10], [13], [20]–[22], [28], [30]–[34], or extracted only concepts [35].

1) Statistical approach: Statistical methods analyse the frequency of terms and their co-occurrence in a document. They tend to be efficient and transportable but imprecise, because semantics of terms is not considered. As they do not depend on specific language or domain, these methods can be used in CMM from documents in different languages and different areas. These methods are commonly used in combination with other methods such as machine learning or linguistic. Created CMs are mostly non-hierarchic, and they can be used as scaffolds for exploration of the source document content.

Commonly used statistical methods are analysis of co-occurrences between terms [10], [11], [21], [28], [35], self-organizing maps [22] and different term frequency analyzing techniques such as TF-IDF [15], [24], [26], [32], LSA [17] and PCA [20].

2) Machine learning: Machine learning methods base its functionality on supervised or unsupervised learning. Learned rules are used for extraction of concepts and relationships from unknown data. Classification, association rules and clustering are techniques mostly used in this process.

Classification technique is used for extraction of key terms in the TEXCOMON [25] software tool. That application uses a simple *Kea* algorithm, based on a naïve Bayes classifier. Lee et al. [33] used the a priori algorithm and association rules for automatic construction of a CM from learners' wrong answers on test questions. As the method creates only association rules based on questions that are not correctly answered, it can miss information based on correctly answered questions. The improved method [31] creates association rules of all questions and gets results that are more appropriate for use in adaptive learning systems.

A number of methods implement fuzzy reasoning and techniques. Concept frame graph [27] is a hybrid algorithm, which uses grammar analysis and fuzzy clustering techniques for creation of a CM knowledge base represented with concept frames. Each concept frame consists of information about concept: name, context, set of synonyms and relationships among other concepts.

Sue et al. [34] used fuzzy association rules to extract predicates from learners' historical testing records in the two-phase CMM method. Bai and Chen [30] simplified and improved the two-phase method for practical use in adaptive courses. Fuzzy association concept mapping method [18] searches for non-explicit links among concepts, using their weights combined with fuzzy heuristic. Similar approach presented in [32], creates CMs from messages posted to online discussion forums.

3) Usage of dictionary: To define concepts and relationships more precisely, in the extraction phase some researchers use ontology and lists of predefined terms as the seed. Given a single term, it is possible to retrieve terms and relationships that most frequently occur with it across a document collection. Extracted words are narrowing the created CM to the chosen domain. List of terms is stored in a dictionary, which contains the basic form of each term. In case of a more complex dictionary, it contains different forms, meanings and relationships between words and phrases.

In a biomedical domain, researchers [13], [21] used the *Medical Subject Headings* (MeSH) [36] and *Unified Medical Language System* (UMLS) [37] ontologies as a dictionary. In another research [10], a dictionary consisted of important terms, their synonyms and metonyms selected by an expert. Other dictionary examples were data from *Computing ontology project* [15], [38], keywords from academic articles [20] and index terms from books [28].

4) Usage of linguistics tools and techniques: Basic statistical and data mining approaches are extensible with lexical or semantic elements, which can be used as additional features in calculations. Numerical technique would make better predictions if it could consider all similar but slightly different expressions, as a same term. The simplest improvement of numerical techniques is normalization of inflected words using lemmatization or stemming.

Lemmatization is a process of determining the proper morphological base form (lemma) of a given word, using vocabulary and morphological analysis. This process usually involves complex and very demanding tasks such as understanding the context and determining the part-of-speech (POS) of words in a sentence. Stemming is a computational process that removes beginnings or endings of words, trying to remove derivational affixes in correct way. Usually, the created root is not identical to the morphological root of the word. That way of normalization is used when it is sufficient to connect related words to the same stem, even if it is not a valid root [6]. It does not demand detailed morphological analysis of a text as lemmatization. Stemming is often used for normalization texts in English and other less inflected languages, because stemming algorithms are usually very simple and efficient.

The problem with linguistic techniques is that they are limited to a specific language, and linguistic resources must exist for used language. The majority of linguistic tools and methods are based on the English language, and researchers

mostly use them in CMM [14], [15], [17], [18], [23]–[27], [32]. A number of them use the *WordNet* for lemmatization, POS tagging and terms disambiguation. The *WordNet* is designed as a repository of the English language, which groups words into synonym clusters, called *synsets*, provides their general definitions, and shows the various semantic relationships among these clusters [39].

Due to lack of language specific resources, very few researchers used this approach for mining texts in other languages: Brazilian Portuguese [12], Chinese [28], Japanese [16] and Slovak [29].

In order to avoid problems of language techniques, in his PhD thesis Richardson [15] described a method for automatic translation of a CM from one language to another. As practical example, he created maps from theses and dissertations in computing in the English language and translated them to Spanish. During a CMM process, syntactic dependencies between noun phrases were recognized using *WordNet's* morphological functions, and translated into CM propositions. Smaller CMs were summarized using simple heuristics, taking top 20 concepts. For summarization of huge CMs with more than 100 concepts, the value of TF-IDF indicator was used. Created CMs were translated into Spanish using an algorithm for translation of individual words and shorter phrases based on bilingual dictionary.

Willis and Miertschin [19] used centering resonance analysis for creation of CMs used in the process of students' assessment. Using linguistics theory, that method creates word network of nouns and noun phrases in order to represent main concepts, their influence, and their interrelationships.

C. Mining from structured textual data sources

Ontologies are mostly used as structured textual data sources in CMM. Their usage in computing has been increased during last decade, especially after promotion of the semantic Web. In the context of knowledge sharing, ontology is a description of objects and relationships between them [40].

CMs and ontologies are quite similar, as ontology can be formalized as a triple $subject - predicate - object$, and a CM as $concept - relation - concept$. Both of them consist of classes or concepts and relationships among them. Unlike CMs, ontologies are more formal and expressive with attributes, values and restrictions. Basic approach used for translation of ontology to a CM is direct mapping of ontology classes and associations into concepts and relationships.

Kim et al. [41] proposed a way for translation of ontology of English vocabulary into a customized CM. The translation is processed by the software agent that directly maps ontology classes and properties to CM propositions. The algorithm described in [42] follows the same approach. The first part of that algorithm searches for ontology class hierarchy and discovers instances of classes. Synonyms, intersections and unions among classes are translated into links between concepts. Equally, all properties, data types and values became new concepts. At the end, the algorithm checks and corrects symmetric and transitive links.

Kumazawa et al. [43] created a method for visualizing data from domain-independent ontology in the form of multiple CMs. Depending on user's interests and perspectives, it allows visualization of more domain-dependent parts of a large ontology. The user chooses a starting concept, and other concepts are extracted from ontology according to *is-a*, *part-of*, and *attribute-of* relationships with the starting concept. The depth of extraction depends on user's choice.

D. Mining from non-text data sources

Less-commonly used sources for CMM are non-text, such as motion stream [44], speech [45] and video [46]. The first phase of CMM from such sources is data transformation into unstructured text. For audio sources, automatic speech recognition methods are based on hidden Markov models (HMM) [47]. Elements in video sources are recognized using computer vision and pattern recognition methods [48]. After unstructured text is created, previously described CMM methods are used.

E. Evaluation of a CMM process

An important problem emphasized in the studies is the need for an objective evaluation method, which will be used for comparison of CMM results. Although most of researchers evaluate created CMs, their evaluation is mostly subjective and based on personal opinion of one or few experts. If we consider only the quality of extracted concepts, authors argue relevance of their extraction methods to 65–90%, and relevance of complete, automatically created maps to 50–80%.

Humans are good as evaluators of created CMs because they know how to capture important terms and produce a well-formed CM. A problem appears because different persons look at the source content in their own way and the process of choosing the key concepts is matter of a personal judgment. The quality of evaluation is hard to control, and multiple evaluators will produce more or less agreed maps. That problem can be reduced introducing an inter-human agreement among skilful human evaluators, who have to be familiar both, with the CMM process and the educational requirements of created CMs. The map agreed on that way, can be used as a gold standard for evaluation [49] of automatically created CMs.

Villalon and Calvo [2] proposed an objective method for evaluation of CMs, based on a gold standard. That method calculates difference between human-created and machine-generated CMs, with reference to an inter-human agreement. By their recommendation, a gold standard used for evaluation should be a set of CMs created by two or more human annotators who identify the relevant propositions. As a distance measure between maps, they recommend quantitative measures such as precision and recall, successfully used in evaluation of created ontologies, and in different NLP areas [49].

F. Comments on concept map mining approaches

Fully automatic production of human-quality CM from a given document is a hard problem, which has not been satisfactorily resolved yet. It is not enough just to extract words from a document, but to find and label relevant concepts and relationships among them. The problem with automatically created CMs is that the number of extracted concepts is often huge or too small, and some concepts are irrelevant to the problem domain. It is hard to find correct complex phrases and labels of links.

A frequent problem that occurs in the recognition of relationships, is finding to which noun phrase a pronoun phrase

refers. Problem is known as anaphora resolution [50]. Without proper anaphora resolution, much of semantic information from a text could be lost. Determining a link label and direction is an additional difficulty.

Most of the proposed CMM methods have been applied and customized for small corpora in a known domain. These techniques mainly learn a language model, or a set of rules from training documents, and then apply that model or rules to new texts. In such closed environment, approaches that use profound linguistic knowledge and techniques can achieve good results. Models learned in this way are effective on documents similar to training documents, but behave quite poorly when applied to documents of a different type. As a result, this approach cannot be efficiently used in an open environment like the Web due to the diversity of problem domains, high usage of computer resources and the cost of creating an equally diverse set of training documents. For use in such environments, researchers should abandon extensive domain and linguistic knowledge, and adopt knowledge-poorer strategies.

IV. Concept map mining from texts in the Croatian language

In this chapter, a CMM procedure for automatic creation of CMs from unstructured texts in the Croatian language is proposed and described in detail.

A. General procedure

This CMM procedure creates CMs from unstructured texts using statistical and data mining techniques, enriched with linguistic tools. A single CM is created from one document. The first iteration is based on statistical and data mining techniques, with very limited usage of linguistic tools. That approach is suitable for mining of texts with no regard to their language. Depending on results, techniques that require deeper linguistic knowledge will be used in the second iteration. Main steps of the procedure are shown in the Fig. 2.

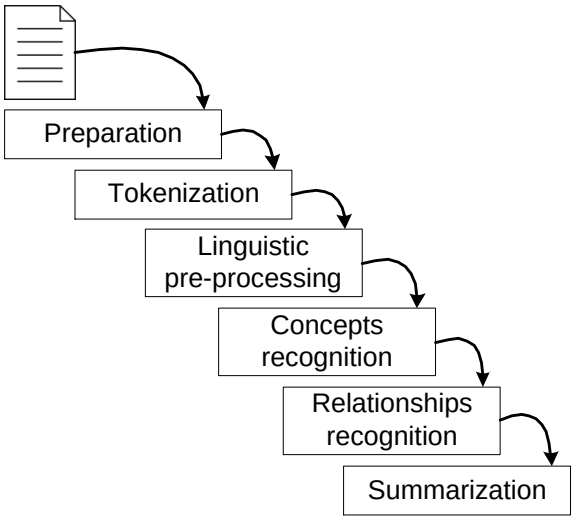

Fig. 2. General procedure for CMM of unstructured textual sources

In the first, preparation phase, the text is retrieved from a chosen source. All elements without information are removed and the cleansed text, enriched with semantic information is stored. In the tokenization phase, the fetched document is divided into sentences, and every sentence into tokens.

The next phase is linguistic pre-processing. From a set of tokens, the stop words without information value are removed. If techniques that require deeper linguistic knowledge are used, every token should be associated with its lexical role using the POS tagger. Anaphora resolution could be performed on POS tagged sentences, and for problematic pronouns, corresponding nouns could be determined. Tokenization of longer documents results in a higher dimensionality of the set, which can be reduced by word normalization using lemmatization or stemming. The result of normalization is a reduced set of terms normalized to its basic form.

Candidates for concepts are chosen from a set of tokens, using previously created dictionary of key terms in chosen domain. If POS tags are determined in the previous phase, they can help in this process because nouns and noun phrases are considered as main concepts candidates.

Candidates for relationships are the words semantically connected to extracted concepts. Usage of POS tagged verb phrases may lead to extraction of less ambiguous relationships. Combining normalized concepts with links, a set of propositions is created. During the final, summarization phase, the most important propositions are chosen from the set by their statistical significance, and they serve as a base for CM usage.

B. Details of the procedure

1) Data source: This research is focused on creation of CMs from texts in the Croatian language. Data source used for analysis will be legal documents in science and education area, publicly available from the Web page of the *Official Gazette of the Republic of Croatia* [51]. It is an accessible source of legal documents written in the correct Croatian language. As legal documents are known to be complex and hard to understand, created CMs could be used as a helpful addition for better and easier understanding of specific document.

The first step of this process is selection of documents suitable for analysis. Primary choices are papers concerning institutions and groups of people, such as laws, regulations, injunctions and rulebooks. Papers focused on individuals on specific positions, such as decisions on appointment or dismissal, are left out because they are very short, written using same phrases, and for those reasons, they a not valuable source of information.

Standard semantic structure of a legal text can be useful in a mining process. Important parts of a document that describe semantics include the author, addressee, type and the title of a document, titles of parts, articles and sections in a document, definition of abbreviations and terms, expiry period and connection to other documents.

HTML documents will be downloaded from the Web page and stored locally. The system should determine a correct code page for each document, because, documents are encoded using different encoding schemes, such as Windows-1250 or UTF-8. An algorithm will check HTML encoding attribute from the heading of each document. If the value of that attribute does not exist, the algorithm will use heuristics trying to recognize the correct code page. In addition, it will decode special character entities, like *&* to give the correct characters.

Text and basic semantic information will be extracted and analysed, and documents that are not suitable for this process

will be abandoned. Suitable documents will be cleaned from mark-up and other data without information value and stored for further processing. All retrieved documents will be divided into two parts: a learning set and a test set. The learning set will be created from $\frac{3}{4}$ of randomly chosen documents and the remaining $\frac{1}{4}$ of the set will be used for the test purpose.

2) Creation of a dictionary: The dictionary of the most important terms in the project area will be created extracting words and phrases from a learning set of documents based on their frequency. Firstly, the stop words will be removed from the set of terms. Stop words are sometimes building elements in multiword phrases, so this extraction should be done after those phrases are recognized. All terms with a similar meaning will be grouped by similarity using statistical analysis such as LSA or probabilistic latent semantic analysis (PLSA).

Extracted terms will be reduced to their basic form and stored in a hash table. The key in hash table is the basic form of a term. As a part of the dictionary, an acronym-mapping table will be created to connect acronyms with mapping phrases. Abbreviations depend on a specific document and they will not be stored in the dictionary, although a list of phrases connected with each abbreviation can be created, and probability of connection can be determined. Those probabilities could be used in later phases for resolving ambiguities of abbreviations' meaning.

3) Tokenization: During the tokenization phase, basic language elements will be identified. These elements are usually words or phrases separated by non-alphanumeric tokens, such as white spaces and punctuation marks. Simple strategy will be used to split sentence to all non-alphanumeric characters.

All words will be converted to lowercase. There will be exceptions, especially for abbreviations, hyphenations, digits and some name entities. Many proper nouns are derived from common nouns and distinguished only by case. To recognize them, a simple heuristic algorithm will be used. This algorithm leaves capitalized mid-sentence words and converts to lowercase the words at the beginning of sentence. These words are usually ordinary words that have to be capitalized because of grammatical rules. An abbreviation-mapping table will be created to map all abbreviations in a document with their mapping phrases.

In creation of a token list, it is important to remember the connection of each word and sentence, because later analyses as a feature use this information. Weight will be assigned to every term, and terms extracted from a document's semantic data will have larger weight than terms from an ordinary text. That weight will be used in the process of summarization of a CM, because according to my presumption, semantic data carry more precise information than ordinary text.

4) Linguistic pre-processing: At the beginning of this phase, stop words will be removed from the set of tokens, and all the words will be normalized. In highly inflected languages like Croatian, this phase is very important, and this process reduces inflectional forms of a word to a common base. Word normalization can be achieved using lemmatization or stemming.

Several studies have examined different methods of stemming [52]–[56] and lemmatization [57]–[59] in the Croatian language. In the first iteration, stemming will be used as a normalization technique. A simple stemmer for the Croatian language will be built using rule-based approach similar to those described in [54]. In the second iteration, depending on the achieved results, techniques that require deeper linguistic knowledge will be used. Lemmatization will be the choice for the normalization technique and during processing, extracted words will be POS tagged.

Initial analysis of a sample text showed that in this process, anaphora would not be a significant problem. The reason lay in the writing style used for legal texts, where authors try to be as precise as possible and they generally avoid the use of pronouns.

5) Concepts extraction: Concepts extraction is a process of discovering potential candidates for concepts in a CM. Generally, the subject of a sentence represents the first concept, and the second concept is represented by the object of a sentence.

The algorithm for extraction is based on the dictionary. A set of rules will be created and all the terms whose base forms are found in the dictionary will be directly marked as concept candidates. In addition, all adjacent words that frequently occur in the neighbourhood will be marked the same, and a connection among them will be established. A word is considered as neighbour if it is part of the same sentence as a concept, and is within a specific distance from a concept. TF-IDF indices will be used for a frequency calculation. In the second iteration of the research, knowledge of a POS role of the chosen term can be a useful addition to this process, as nouns and noun phrases are main candidates for concepts.

6) Relationships extraction: Concepts in a CM are semantically related and during this phase, links between them will be recognized. A type and a label of relationship between two concepts in a simple sentence can be identified by the main verb in the sentence. For each pair of concept candidates, all words positioned in their neighbourhood will be temporarily saved as candidates for a link. A set of rules will be created for extraction of link candidates from that temporary storage, based on frequency of their appearance in the concepts neighbourhood.

If the frequency of appearance of two concepts in the same sentence is high, then the degree of relation between them is high. If the frequency of appearance of two concepts in the neighbouring sentences is high, then a degree of relation between them is medium. Frequency will be calculated using the value of TF-IDF indices. Each relation of high or medium degree will be saved as a three-element set: two adjacent concepts, label of the link between them and the strength of that relation. Each created set presents one proposition, and it will be used as input in the next, summarization phase. In the second iteration of this research, POS tags can be used, as verbs and adverbs are main candidates for relationships.

7) Concept map summarization and generation: The CM that provides an overview of the document's contents, with minimal redundancy is result of the summarization phase. In the first iteration, statistical techniques used for summarization of dictionary terms will be used on propositions created in the previous phase. Relative importance of propositions within a document will be calculated, and top propositions will be extracted. Propositions with higher calculated values will be positioned higher in the CM hierarchy, and the strongest proposition will be marked as a starting one. Based on the comparison of summarization results in the first iteration of

research, specific technique will be chosen for use in the second iteration.

Optimal number of concepts in the final CM will be determined and summarized propositions will be stored in the CXL format [60] based on XML. Maps in that format can be visualized using different concept mapping tools.

8) Evaluation of created concept maps: Evaluation of automatically created CMs will be done using CMs created by human annotators as a gold standard. Two to four human annotators will create CMs from randomly chosen documents in a test set. From those maps, propositions will be extracted and used for an evaluation. Inter-human agreement among created documents will be measured using kappa statistics or another method that measures the extent to which annotators agree in their interpretations [61].

Depending on the calculated value, annotation process will be tuned. Automatically created CM will be compared with a gold standard map of the same document using precision and recall measures, commonly used for evaluation in the NLP area.

V. Conclusion

This paper deals with CMM methods and approaches. In the initial part, a definition and a literature review of that area are given. In the following chapter, a procedure for CMM from unstructured texts is proposed and described in detail.

The described method is focused on CMM from texts in the Croatian language, which is a highly inflective language. Most of the other similar research was dealing with low or moderately inflective languages such as English, Portuguese, Japanese or Chinese.

This research uses legal texts in the area of science and education as a source of data. These documents vary in size from very short to very long. Other studies are focused on mining texts of similar size - mostly shorter, such as essays, newspaper articles or academic papers, and in some studies longer, such as theses and dissertations.

The proposed CMM method uses unstructured text of a document and semantic information embedded in the document, such as information on author, addressee, type and titles of parts of the document and information of connection between that document and other documents. Semantic information extracted from that data will have a bigger weight than an ordinary text. To my best knowledge, this approach, which is usual for summarization techniques, has not been used as a part of CMM methods described in other studies.

The next step of this research is practical implementation of the proposed procedure. Research will be conducted in two iterations. In the first iteration, "linguistically poor" approach with statistical and data mining methods will be used. For creation of a prototype in that phase, existing programming tools and applications will be used as much as possible. At the end of that iteration, a test of prototype and an evaluation of created maps will be conducted. Depending on the results, mining and evaluation procedures will be adjusted.

One of goals that I will try to achieve in this research is to create a system, which is applicable as a real-time operation. That means that precision may be sacrificed for a better and faster response. Considering this, I shall have to make a decision about usage of richer linguistic tools and techniques, to be used in later stages of the research. If tools chosen for a specific task cannot achieve expected results fully, I shall modify or rebuild them.

In the second iteration of this research, a modified procedure and tools will be used for creation and evaluation of the second version of the prototype. In the final part of the research, created features will be integrated in the *SelfTaught* [62] personal learning environment application.

References

[1] J. D. Novak and A. J. Cañas, "The theory underlying concept maps and how to construct and use them," IHMC CmapTools 2006-01, Rev 01-2008, Florida Institute for Human and Machine Cognition, Pensacola Fl, Tech. Rep., January 2008.

[2] J. J. Villalon and R. A. Calvo, "Concept map mining: A definition and a framework for its evaluation," *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, vol. 3, pp. 357–360, 2008.

[3] B. M. Kramer and J. Mylopoulos, "Representation, knowledge," in *Encyclopedia Of Artificial Intelligence*. Wiley-Interscience Publication, 1987, vol. 2, pp. 206–214.

[4] M. D. McNeese, B. S. Zaff, K. J. Peio, D. E. Snyder, J. C. Duncan, and M. R. McFarren, "An advanced knowledge and design acquisition methodology for the pilot's associate," Harry C. Armstrong Aerospace Medical Research Laboratory, Human Systems Division, Tech. Rep., 1990.

[5] W. M. K. Trochim, "An introduction to concept mapping for planning and evaluation," *Evaluation and Program Planning*, vol. 12, no. 1, pp. 1–16, January 1989.

[6] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge University Press, 1999.

[7] D. Das and A. F. T. Martins, "A survey on automatic text summarization," November 2007. [Online]. Available: http://www.cs.cmu.edu/~afm/Home_files/Das_Martins_survey_summarization.pdf

[8] K. Spärck Jones, "Automatic summarising: The state of the art," *Inf. Process. Manage.*, vol. 43, pp. 1449–1481, November 2007.

[9] M. S. Binwahlan, N. Salim, and L. Suanmali, "Fuzzy swarm based text summarization," *Journal of Computer Science*, vol. 5, no. 5, pp. 338–346, 2009.

[10] R. B. Clariana and R. Koul, "A computer-based approach for translating text into concept map-like representations," in *Proceedings of the First International Conference on Concept Mapping*, Pamplona, Spain, September 2004, pp. 131–134.

[11] B. Gaines and M. L. G. Shaw, "Using knowledge acquisition and representation tools to support scientific communities," in *Proceedings of the 12^{th} National Conference on Artificial Intelligence (AAAI94)*. Menlo Park: AAAI Press/MIT Press, 1994, pp. 707–714.

[12] J. H. Kowata, D. Cury, and M. C. S. Boeres, "Concept maps core elements candidates recognition from text," in *Proceedings of Fourth International Conference on Concept Mapping*, Viña del Mar, Chile, 2010, pp. 120–127.

[13] P. Matykiewicz, W. Duch, and J. Pestian, "Nonambiguous concept mapping in medical domain," in *ICAISC*, 2006, pp. 941–950.

[14] A. Oliveira, F. C. Pereira, and A. Cardoso, "Automatic reading and learning from text," in *Proceedings of the International Symp. on Artificial Intelligence*, 2002, pp. 1–12.

[15] R. Richardson, "Using concept maps as a tool for cross-language relevance determination," Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2007.

[16] H. Saito, T. Ohno, F. Ozaki, K. Saito, T. Maeda, and A. Ohuchi, "A semi-automatic construction method of concept map based on dialog contents," in *International Conference on Computers in Education*, 2001.

[17] J. Villalon and R. A. Calvo, "Concept extraction from student essays, towards concept map mining," in *Proceedings of the Ninth IEEE International Conference on Advanced Learning Technologies*, ser. ICALT '09, Washington, DC, USA, 2009, pp. 221–225.

[18] W. Wang, C. Cheung, W. Lee, and S. Kwok, "Mining knowledge from natural language texts using fuzzy associated concept mapping," *Information Processing & Management*, vol. 44, no. 5, pp. 1707–1719, 2008.

[19] C. L. Willis and S. L. Miertschin, "Centering resonance analysis: a potential tool for IT program assessment," in *Proceedings of the 2010 ACM conference on Information technology education*, ser. SIGITE '10. New York, NY, USA: ACM, 2010, pp. 135–142.

[20] N.-S. Chen, Kinshuk, C.-W. Wei, and H.-J. Chen, "Mining e-learning domain concept map from academic articles," *Comput. Educ.*, vol. 50, pp. 1009–1021, April 2008.

[21] J. W. Cooper, "Visualization of relational text information for biomedical knowledge discovery," in *Information Visualization Interfaces for Retrieval and Analysis workshop*. ACM SIGIR, 2003.

[22] M. Hagiwara, "Self-organizing concept maps," in *Proceedings of 1995 IEEE International Conference on Systems Man and Cybernetics Intelligent Systems for the 21^{st} Century*, vol. 1. IEEE, 1995, pp. 447–451.

[23] L. Kof, R. Gacitua, M. Rouncefield, and P. Sawyer, "Concept mapping as a means of requirements tracing," in *Managing Requirements Knowledge (MARK), 2010 Third International Workshop on*, September 2010, pp. 22–31.

[24] A. Valerio and D. Leake, "Jump-starting concept map construction with knowledge extracted from documents," in *Proceedings of the Second International Conference on Concept Mapping*, San José, Costa Rica, 2006, pp. 181–188.

[25] A. Zouaq and R. Nkambou, "Building domain ontologies from text for educational purposes," *IEEE Trans. Learn. Technol.*, vol. 1, pp. 49–62, January 2008.

[26] A. Zouaq, D. Gasevic, and M. Hatala, "Ontologizing concept maps using graph theory," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, ser. SAC '11. New York, NY, USA: ACM, 2011, pp. 1687–1692.

[27] K. Rajaraman and A.-H. Tan, "Knowledge discovery from texts: a concept frame graph approach," in *Proceedings of the 11^{th} International Conference on Information and Knowledge Management*, ser. CIKM '02. New York, NY, USA: ACM, 2002, pp. 669–671.

[28] Y.-H. Tseng, C.-Y. Chang, S.-N. C. Rundgren, and C.-J. Rundgren, "Mining concept maps from news stories for measuring civic scientific literacy in media," *Comput. Educ.*, vol. 55, pp. 165–177, August 2010.

[29] K. Furdík, J. Paralič, and S. P., "Classification and automatic concept map creation in elearning environment," in *Proceedings of the Czech-Slovak scientific conference Znalosti*, 2008, pp. 78–89.

[30] S.-M. Bai and S.-M. Chen, "Automatically constructing concept maps based on fuzzy rules for adapting learning systems," *Expert Syst. Appl.*, vol. 35, pp. 41–49, July 2008.

[31] S.-M. Chen and S.-M. Bai, "Using data mining techniques to automatically construct concept maps for adaptive learning systems," *Expert Syst. Appl.*, vol. 37, pp. 4496–4503, June 2010.

[32] R. Y. K. Lau, D. Song, Y. Li, T. C. H. Cheung, and J.-X. Hao, "Toward a fuzzy domain ontology extraction method for adaptive e-learning," *IEEE Trans. on Knowl. and Data Eng.*, vol. 21, pp. 800–813, June 2009.

[33] C.-H. Lee, G.-G. Lee, and Y. Leu, "Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning," *Expert Syst. Appl.*, vol. 36, pp. 1675–1684, March 2009.

[34] P.-C. Sue, J.-F. Weng, J.-M. Su, and S.-S. Tseng, "A new approach for constructing the concept map," *IEEE International Conference on Advanced Learning Technologies*, vol. 49, no. 3, pp. 76–80, 2004.

[35] A. J. Cañas, M. Carvalho, M. Arguedas, D. B. Leake, A. Maguitman, and T. Reichherzer, "Mining the web to suggest concepts during concep map construction," in *Proceedings of the First International Conference on Concept Mapping*, Pamplona, Spain, September 2004, pp. 135–142.

[36] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bull Med Libr Assoc.*, vol. 88, no. 3, pp. 265–266, July 2000.

[37] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.

[38] L. N. Cassel, G. Davies, W. Fone, A. Hacquebard, J. Impagliazzo, R. LeBlanc, J. C. Little, A. McGettrick, and M. Pedrona, "The computing ontology: application in education," *SIGCSE Bull.*, vol. 39, pp. 171–183, December 2007.

[39] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, pp. 39–41, November 1995.

[40] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, pp. 199–220, June 1993.

[41] K. Kim, C. M. Kim, and S. B. Kim, "Building/visualizing an individualized English vocabulary ontology," in *Proceedings of the 5th WSEAS international conference on Applied computer science*, ser. ACOS'06. Stevens Point, Wisconsin, USA: WSEAS, 2006, pp. 258–263.

[42] V. Graudina and J. Grundspenkis, "Concept map generation from OWL ontologies," in *Proceedings of the Third International Conference on Concept Mapping*, Tallinn, Estonia and Helsinki, Finland, 2008, pp. 263–270.

[43] T. Kumazawa, O. Saito, K. Kozaki, T. Matsui, and R. Mizoguchi, "Toward knowledge structuring of sustainability science based on ontology engineering," *Sustainability Science*, vol. 4, no. 1, pp. 99–116, 2009.

[44] Y. Yang, H. Leung, L. Yue, and L. Deng, "Automatically constructing a compact concept map of dance motion with motion captured data," in *Advances in Web-Based Learning - ICWL 2010*. Springer Berlin / Heidelberg, 2010, vol. 6483, pp. 329–338.

[45] K. Böhm and L. Maicher, "Real-time generation of topic maps from speech streams," in *Charting the Topic Maps Research and Applications Landscape*. Springer Berlin / Heidelberg, 2006, vol. 3873, pp. 112–124.

[46] A. Badii, C. Lallah, M. Zhu, and M. Crouch, "Semi-automatic knowledge extraction, representation, and context-sensitive intelligent retrieval of video content using collateral context modelling with scalable ontological networks," in *Multimedia Analysis, Processing and Communications*. Springer Berlin / Heidelberg, 2011, vol. 346, pp. 459–474.

[47] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Speech Recognition, Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[48] N. Sebe and M. S. Lew, *Robust Computer Vision Theory and Applications*, ser. Computational Imaging and Vision, M. A. Viergever, Ed. Kluwer Academic Publishers, 2003, vol. 26.

[49] K. Dellschaft and S. Staab, "On how to perform a gold standard based evaluation of ontology learning," *Learning*, vol. 4273, pp. 228–241, 2006.

[50] R. Mitkov, "Outstanding issues in anaphora resolution," in *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, ser. CICLing '01. London, UK: Springer-Verlag, 2001, pp. 110–125.

[51] Narodne novine. (2011, August) Official Gazette of the Republic of Croatia. In Croatian. [Online]. Available: http://narodne-novine.nn.hr/

[52] D. Kalpić, "Automated coding of census data," *Journal of Official Statistics*, vol. 10, no. 4, pp. 449–463, 1994.

[53] D. Lauc, T. Lauc, D. Boras, and S. Ristov, "Developing text retrieval system using robust morphological parsing," in *Proceedings of 20^{th} International Conference on Information Technology Interfaces*, 1998, pp. 61–65.

[54] N. Ljubešić, D. Boras, and O. Kubelka, "Retrieving information in Croatian: Building a simple and efficient rule-based stemmer," in *Digital Information and Heritage*. Zagreb: Department for Information Sciences, Faculty of Humanities and Social Sciences, 2007, pp. 313–320.

[55] J. Snajder, B. Dalbelo Basic, and M. Tadic, "Automatic acquisition of inflectional lexica for morphological normalisation," *Information Processing & Management*, vol. 44, no. 5, pp. 1720–1731, 2008.

[56] J. Šnajder and B. Dalbelo Bašić, "String distance-based stemming of the highly inflected Croatian language," in *Proceedings of the International Conference RANLP-2009*. Borovets, Bulgaria: Association for Computational Linguistics, September 2009, pp. 411–415.

[57] M. Tadić and S. Fulgosi, "Building the Croatian morphological lexicon," in *Proceedings of EACL*, 2003, pp. 41–46.

[58] R. Lujo, "Locating similar logical units in textual documents in Croatian language," Master's thesis, Faculty of electrical engineering and computing, 2010, in Croatian.

[59] J. Šnajder, "Morphological normalization of texts in croatian language for text mining and information retrieval," Ph.D. dissertation, Faculty of electrical engineering and computing, 2010, in Croatian.

[60] A. J. Cañas, G. Hill, L. Bunch, R. Carff, T. Eskridge, and C. Peréz, "KEA: A knowledge exchange architecture based on web services, concept maps and cmaptools," in *Proceedings of the Second International Conference on Concept Mapping*, San José, Costa Rica, 2006, pp. 304–310.

[61] K. L. Gwet, *Handbook of inter-rater reliability*, 2nd ed. Advanced Analytics, LLC, 2010.

[62] K. Žubrinić, "Software for creation of personal learning environment," Master's thesis, Faculty of electrical engineering and computing, Zagreb, 2010, in Croatian.