Innovating with Concept Mapping
Proc. of the Seventh Int. Conference on Concept Mapping
Tallinn, Estonia 2016

AUTOMATIC CONSTRUCTION OF CONCEPT MAPS FROM TEXTS

Camila Z. Aguiar & Davidson Cury, Federal University of Espírito Santo, Brazil
Amal Zouaq, University of Ottawa, Canada
E-mail:camila.zacche.aguiar; dedecury; amal.zouaq @gmail.com

Abstract. The automatic construction of concept maps from texts is still an ongoing research topic, especially when the map should represent, by means of concepts and relationships, a summarization of a complex text. We propose a new method for the automatic generation of concept maps. Based on an analysis of the limitations and best features of the state-of-the-art, this paper introduces a pipeline that comprises (a) grammar rules and depth-first search for the extraction of concepts and links between them from text; (b) anaphora resolution; (c) exploitation of named entities for concept identification; and (d) concept relevance calculation through frequency of occurrence and an analysis of the topology of the concept map. The approach and preliminary results are presented in the form of a prototype, which is still under development.

1 Introduction

According to Novak & Cañas (2008), concepts and relationships form the basis for learning and, therefore, conceptual maps have been extensively used in different situations and for different purposes in education. The standard procedure for building a concept map involves defining a topic or focal question, identifying and listing the most important or "general" concepts related to the topic, ordering the concepts in terms of importance and adding and labeling connecting phrases between concepts. The manual construction of a map requires a significant time and a committed effort in identifying and structuring knowledge, especially when the construction of the map is performed from scratch, that is, when its constituent elements are not predetermined and must be completely identified in some data source. In order to facilitate the process of building concept maps, especially maps in the Novakian style, various technological approaches have been proposed to help automate that process (Richardson & Fox, 2007; Villalon & Calvo, 2011).

The automatic construction of a concept map requires great technological and processing effort. The information extraction techniques must be able to identify concepts that are relevant to a particular domain, identify linking phrases that make the relationship between two concepts significant, define the hierarchy of concepts that will be displayed on the map, and build links between concepts that are not directly evident. Because of these difficulties, many approaches adopt semi-automatic techniques to ensure greater precision in the identification of concepts. In this scenario, we observed (1) the use of predefined list of concepts as the basis for identifying concepts belonging to a particular domain (Clariana & Koul, 2004), (2) the use of ontologies (Graudina & Grundspenkis, 2008) and (3) the use of online documents or concept maps for the extraction of the knowledge related to the domain, in order to assist the process of building or enriching an already available map (Cañas et al., 2004; Valerio & Leake, 2006).

Educational concept maps have been extensively used as learning resources, means of evaluation, instructional organizations, cognitive representations, and for knowledge elicitation and sharing. In addition, concept maps can be used as a summary for large documents (Richardson & Fox, 2007). We are interested in the use of maps in this context, where a complex text can be summarized by a concept map containing only concepts and relationships that represent succinctly what was expressed in a more complex way.

In this paper, we present our prototype based on a new technological approach for the automatic generation of concept maps from scientific texts in English (Aguiar *et al.*, 2015b). The approach will be integrated with a service-based platform, CMPaaS (Cury *et al*., 2014), under development in our laboratory. This platform aims at expanding and integrating basic services such as edition, management, and manipulation of concept maps. These services will be available to anyone in the world. To date, this platform offers services for merging concept maps (Vassoler *et al*., 2015), information retrieval on maps from questions (Perin *et al.*, 2014), and shallow ontologies construction from maps (Pinotte *et al.*, 2015). In this context, the present study proposes a new service on the CMPaaS platform.

This paper is structured as follows: Section 2 presents some related work; in Section 3 we discuss a new architecture for a *Concept Maps Miner*; Section 4 introduces the process for the automatic construction of concept maps; Section 5 presents the experiments and results obtained so far, with the implementation of a prototype; and finally, Section 6 discusses some preliminary considerations.

2 Related Works

A number of studies have focused on the automatic construction of concept maps, or similar representations, for various applications. However, in this paper, we are interested in approaches related to the construction of maps from documents that adopt the following requirements:

- Approaches that use only one data source, that is, one unstructured text in English. We do not limit the requirement to a specific text size (small, regular or long), although the text size will affect the process and outcome of the approach;
- Approaches which adopt semi or completely automatic extraction techniques;
- Approaches that generate maps in the Novakian style, i.e., we consider those which follow the strict definition established by Novak with regard to adopting a focal question, a hierarchy, labels of concepts and relationships, and establishing links and cross-links among concepts.

The following is a summary and discussion of the related works found in this context. Since the approaches presented aimed at the same objective, which is the automatic extraction of concept maps, what differs is the process carried out for the extraction of text elements and the visual result displayed on the generated map (Figure 1).

Valerio & Lake (2006) introduces the segmentation of web document in topics, the generation of a map for each topic and then the fusion of all maps. Their approach uses morphological and syntactic analysis and relies on the standardization of terms with synonyms and stemming. It identifies noun phrases as concepts (represented as multi-words when applicable) in the parsing tree and applies a statistical analysis for ranking concepts. Triples are defined for each pair of concepts of the tree that has a verbal type dependency. Looking at the map generated by the approach, in Figure 1 (a), we observe: (1) isolated concepts without relationships; (2) map portions fragmented for each sentence; (3) a lack of anaphora resolution; and (4) Propositions with only a direct link in the sentence.

Richardson & Fox (2007) show the generation of maps for a dissertation, containing only the chapters or the overall information of each chapter. Their approach combines morphological and syntactic analysis, anaphora resolution and entity recognition for building semantic primitives with the help of an ontology. Looking at the map generated by the approach, in Figure 1 (b), we observe that the concepts do not seem to represent the relevant information contained in the chapter of a dissertation, therefore preventing a full analysis.

Wang et al. (2008) generate maps from text abstracts. This approach uses morphological and syntactic analysis, identifying the elements based on the structure of the phrases and syntactic rules. It applies normalization to correct orthographic mistakes, and relies on synonyms detection and anaphora resolution. It uses statistical analysis to check the relevance of the propositions. Uncertain propositions are defined by means of user interaction through questions. Looking at the map generated by the approach, in Figure 1 (c), we observe: (1) map is fragmented in portions; (2) the approach assigns very long labels to concepts; (3) it accepts pronouns as labels for concepts; (4) it accepts prepositions as labels for relationships.

Villalon & Calvo (2011) use linguistic techniques to generate maps for descriptive texts made by students. Their approach creates a dependency parse tree which is transformed into a terminological map. The concepts are extracted from the vertices and the relationships based on the shortest path between two concepts. It applies a set of semantic rules for the terminological map reduction. Besides, it performs a statistical analysis for the computation of terms' relevance. Analyzing the map generated by the approach, in Figure 1 (d), we observe that (1) the map is fragmented in portions and (2) it accepts prepositions as label for the relationships.

Qasim et al. (2013) uses language dependencies to generate maps for scientific articles. It explores the graph of dependencies to extract concepts and relationships (e.g. hyponyms) based on the type of dependencies in the graph. It performs a statistical analysis for the identification of terms' relevance. It uses similarity metrics and clustering to group concepts based on their relationships. Triples are defined if the concepts belong to the same sentence and if the terms found are of the form <subject-verb-object>. Analyzing the map generated by the approach, in Figure 1 (e), we observe that the labels of relationships are defined by the assigned type of dependence. The displayed map is generated from a single sentence.

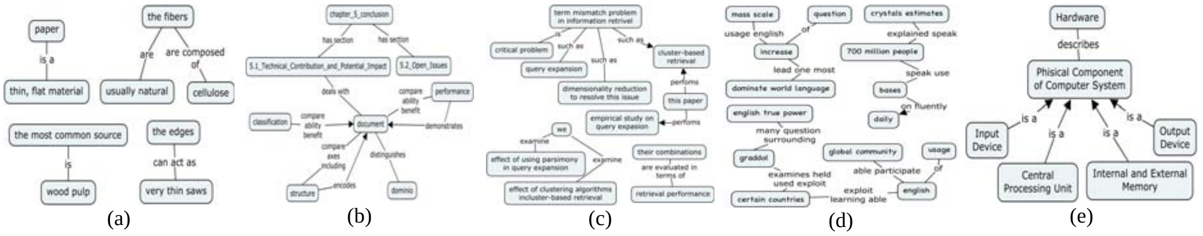


Figure 1. Maps generated by the approaches discussed in Section 2.

According to the analysis carried out on the approaches, we can identify the following challenges: (1) a difficulty in defining small and meaningful labels; (2) A challenge for the identification of relevant concepts related to a given domain; (3) Difficulties for establishing links between concepts which are not directly evidenced in the text; and (4) The complexity of the identification of the domain of a document.

3 A Proposal for a Concept Map Miner (CMM)

A Concept Map Miner is defined as a process for the automatic extraction of concept maps from documents targeted to the educational context (Villalón & Calvo, 2011). This can be formalized by defining a document D as a set D = {Cd, Rd}, where Cd is the set of all concepts, and Rd the set of all links extracted from the document. This extraction process may be synthesized in the following steps: Concept Identification, Relationship Identification and Summarization. The result is a concept map CM = {C, R, T}, where CM is defined by the set of concepts C, relationships R and a topological organization T, as shown in Figure 2 (a).

Using a different perspective from the one by Villalón & Calvo (2011), we propose a process for building maps covering four axes of interest: the *Data Source Description* step defines the techniques used in the information extraction process as well as the appropriate manipulation methods such as the linguistic, statistical, and machine learning techniques, the element retrieval and identification (detailed in Aguiar *et al*., 2016a); the *Domain Definition* step identifies the relevant concepts of the domain; the *Elements Identification* step can be regarded as the core of the process, which makes use of the earlier steps to extract concepts and relationships; and the *Map Visualization* step specifies the graphic positioning of propositions in the concept map, since maps are graphical tools and knowledge visualization is part of the learning process. Such axes should be understood as the steps for the automatic extraction of concept maps from texts (Aguiar *et al*, 2015b).

The proposed process is presented in Figure 2. Figure 2 (b) starts by the description of the data source to characterize a document *D*.

A document *D* of size *n* can be defined as

$D = \{d_1 \ldots d_n\}$ where

d_i, , i=1…, n is a term in D.

A set of concepts can be defined as

$C = \{c_1 \ldots c_n\}$ where

$C \subseteq D$

and

c_i, is a term d_i that represents a concept or entity for the domain.

A set of relationships can be defined as

$R = \{r_1 \ldots r_n\}$ where

$R \subseteq D$

and

r_i is a set of concatenated terms d1_..._d_i that represent a relation between concepts.

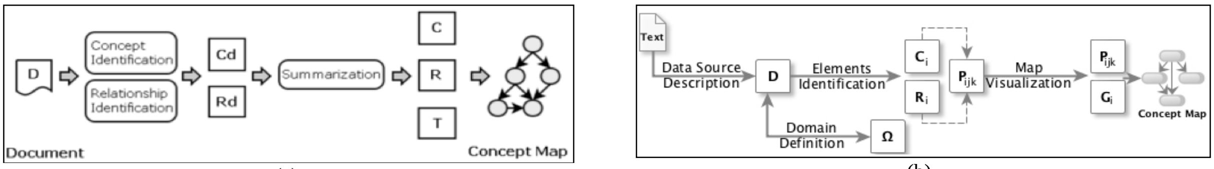
The document *D* is used as an input to the Domain Definition step for the discovery of the document domain Ω. The domain Ω is the union of concepts *C* extracted from a document *D*.

A proposition can be defined as

$P_{ijk} = \{c_i, r_j, c_k\}$ where

$c_i \in C$ and $c_k \in C$ and $r_j \in R$.

During the Map Visualization step, for each proposition P_{ijk}, we assign a graphical position G_i to form a set of propositions organized topologically in the concept map defined as $CM = \{P_{ijk}, G_i\}$.



(a) (b)

Figure 2. The Process of the Concept Map Miner.

The description of the data source impacts the whole process of building the conceptual map. In this context, we identified several variations in the approaches used for the extraction of concept maps in terms of size and quantity of information available in the data source. (1) For the size, we can characterize unstructured data sources as (a) small: small content such as abstracts; (b) regular: few data pages such as academic articles, reports, newspaper, articles etc.; (c) long: extensive data containing a lot of information such as theses and dissertations. (2) According to size dimension, the data source may be represented in two groups: (a) approaches that use a set of documents to represent the knowledge of a domain and (b) approaches that use a single document that represents the knowledge specific to one author.

We believe that one of the challenges for the automatic construction of concept maps from texts is the definition of the domain, i.e. of the text domain or the concepts belonging to the domain. In this scenario, we note the use of semi-automatic techniques where the author identifies the domain of the data source by (1) choosing a suitable ontology (Graudina & Grundspenkis, 2008), or (2) using multiple maps (Valerio et al., 2008), or (3) using a list of concepts (Clariana & Koul, 2004) or (4) by means of a set of documents (Lau et al., 2008).

The identification of the elements, defined as the core of the process, is to extract propositions i.e., *concept-relationship-concept* triples, which will compose the concept map. For a map to be representative, the information must be relevant to the domain, be properly labeled and significantly connected. We have observed approaches that generate fragmented maps with disconnected concepts (Villalón & Calvo, 2011; Valerio *et al*, 2008), or that attribute incomplete or extensive labels (Wang *et al*., 2008), or approaches that fail to create relationships between some concepts (Valerio *et al*, 2008; Villalón & Calvo, 2011), and that do not identify the available linking phrases (Clariana & Koul, 2004).

The map visualization shows the topological structure of propositions identified in the Identification Elements step by means of a graphical interface. In this case, we observed that many approaches are not focused on the map view. Instead, they use third-party tools for these purposes (Villalón & Calvo, 2011; Clariana & Koul, 2004). However, some approaches develop their own display interface including features that facilitate learning, such as (1) a list of occurrences of the concept within the context (Zouaq *et al*., 2007), (2) a partial map view from the perspective of a concept (Lau *et al*., 2008) or (3) the display of the path of a specific concept until the focus question (Kumazawa, 2009).

We believe that the steps proposed by Villalón & Calvo (2011) are embedded in the last two steps of our process, which also follow the three principles of educational utility, simplicity, and subjectivity in the automatic construction of concept maps.

4 On the Approach

Our approach began with the categorization we defined in Aguiar *et al*. (2015a) which, together with the CMM process (Section 3) were used for the development of the approach presented in this paper. The approach is still under implementation as a Web service in Java in conjunction with the framework Spring. The following sections detail each step of the process.

4.1 Description of Data Source

With respect to the Data Source Description step, our approach is characterized by the use of unstructured text in English, in the forms of academic articles whose size is classified as regular according to our definition in Section 3. Being an academic text, the language is formal and the text quality is ensured by its publication. The data source is limited to a single document, since the goal is the representation of the knowledge extracted from text as a summary. The data source is chosen by the User and used as input to the whole process.

4.2 Definition of the Domain

With respect to the Domain Definition step, our approach is characterized by relying on a human User, who initially chooses the data source in step 4.1, to assist the process of domain identification. For the semi automatic Domain Identification of a given text, our approach proposes the use of supervised learning techniques of clustering and classification but this is left for future work. For the identification of the elements belonging to a particular domain, we propose the use of an automatically-built domain thesaurus at this stage. The thesaurus is represents a conceptual vocabulary for a given domain and is built incrementally as new texts from the same domain are processed to build concept maps.

4.3 Identification of the Elements

For the Identification of Elements (concepts, linking phrases) we defined the following rules: (1) labels in concepts are formed by nouns and linking phrases contain a verb; (2) the generated propositions should represent the information contained in the data source; (3) the generated propositions must belong to the identified domain; (4) propositions belonging to the same domain should be connected by linking phrases.

The process for the identification of the elements, shown in Figure 3, is divided into 12 steps, starting from the Preparation step with the reception of the data source, and ending at the Ranking and Summarization step with the construction of propositions of the form *concept-relationship-concept*. Most of these steps are implemented using Stanford CoreNLP (Manning *at al.,* 2014), a set of tools for analyzing natural language provided by Stanford University, together with algorithms developed in our approach.

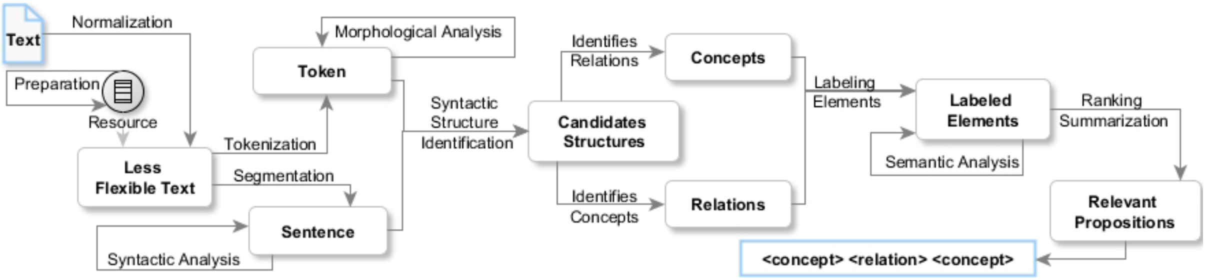


Figure 3. Process adopted by our approach.

The **Preparation** step is responsible for providing all the resources needed for performing the extraction process. In this case it includes the manual mapping of prepositions with a predefined linking phrase including the preposition. For example, the preposition "between" is mapped to "appear between". This step is performed only once.

The **Normalization** step modifies the data source to ease further processing. This comprises (a) eliminating label markers, tags and font style; (b) removing special characters; (c) solving genitive forms by converting possessive case to a normal form; (d) removing non-propositional sentences; and (e) resolving anaphora. Stop words are not removed in this step, since their removal negatively influences the later processes.

The Steps for **Tokenization** and **Morphological Analysis** are performed in parallel, focusing on individual terms. The first step divides the text into tokens; the second identifies the part of speech of each identified token. We use the tokens that contain the tag (a) Noun (NN) and its related Plural Noun (NNS), Singular Proper Noun (NNP) and Plural Proper Noun (NNPS); (b) Adjective (JJ); (c) Cardinal Number (CD); (d)Verb (VB) and its derived Past Tense Verb (VBD), Gerund or Present Participle Verb (VBG), Past Participle Verb (VBN), Non-3rd Person Singular Present Verb (VBP), 3rd Person Singular Present Verb (VBZ) and Modal (MD); (e) Determinant (DT); (f) Adverb (RB) and (g)Preposition (IN) and its derived To (TO). Related tags and optional tags are represented in the article by their first tag followed by apostrophe and by the symbol *, respectively. For example, if we want to represent the tag NN and all its related tags, we use NN', or if it is optional we use NN*.

The **Text Segmentation** and **Syntactic Analysis** steps are performed in parallel, focusing on sentences. The first step provides the segmentation of plain text into individual sentences; the second analyses these sentences to build the sentence parse tree. For this we use two types of grammatical analysis: Clause Level Syntagms, representing simple declarative sentences (S) and Phrase Level Syntagms, containing nominal phrases (NP), verbal phrases (VP) and prepositional phrases (PP). Syntagms not included in these rules are removed from the parsing tree.

The **Syntactic Structure Identification** step applies a new segmentation to the parsing tree of each sentence to identify the candidate structures for a proposition. Namely, all the candidate structures which are occurrences of the following pattern (NP (JJ*)(NN’)) ({VP (VB’) | PP (IN’)} (RB’*) (NP (JJ*)(NN’))) are extracted. We use a depth-first search to look for this syntactic pattern, regardless of the intermediate structures between the syntagms. The pattern will be matched only if the syntagms are completely identified, i.e. we define a complete (a) NP syntagm when it contains a core NN or one of its derivatives, (b) VP syntagm when it contains a core VB or one of its derivatives and a complete NP syntagm, (c) PP syntagm when it contains a core IN or a derivative and a complete NP syntagm.

Complete NPs and VPs syntagms are extracted by finding their sub-trees. For searching the occurrences of a pattern, we adopted three general ideas: (a) searching simple declarative syntagms, (b) searching nominal syntagms and (c) searching preposition syntagms.

During the **Concepts** and **Relationships Identification** step, morphological rules are applied to the candidate structures to identify concepts and relationship. For this, we use the following rules: (1) if the candidate structure is composed by a NP syntagm and its nucleus belongs to {JJ, NN’, CD}, then a Concept is identified; (2) if the candidate structure is composed by a VP syntagm and its nucleus belongs to {VB’, IN’, RB}, then a Relationship is identified; (3) if there is a relationship of specialization between two concepts, then the label "is type of" is assigned.

The steps **Labeling Elements** and **Semantic Analysis** are executed together to reduce the ambiguity and improve the labeling of the extracted elements (concepts and relationship). For this, we used the following rules: (a) multi-words are identified by applying grammar rules on the extracted syntagms; (b) similar concepts are assigned a unique label. We use similarity measure LIN (Lin, 1998) to calculate the similarity of two lemmatized concepts through the dictionary lexicon Wordnet (George, 1995). For example, the concept “ability” and “creative_thinking” is considered similar with 0.85 score; (c) prepositions are transformed into preposition phrases according to the mapping defined in the Preparation step; (d) named entities (places, organizations and proper names) are not directly extracted, since we understand that they are instances of classes (concepts) and should not be represented on the map. However, identifying their types can be of interest to our concept map. To this end, named entity detection is performed using Stanford NER[1] and each sentence of the data source that contains it is retained in a textual summary. A query containing the entity label and the entity type is executed on DBPedia. For example, for the concept “David Ausubel” a query is created with rdf:type dbo:Person and foaf:name “David Ausubel”. The abstract result of DBPedia is stored as a vector and compared with the vector representation of the extracted summary through cosine similarity. If the similarity is high, the description is assigned as concept; otherwise, the entity type is assigned. An example of a SPARQL query is shown below where variable “*var*” is replaced by a concept, for example “David Ausubel”.

```
SELECT DISTINCT ?node ?name ?abstract ?descriptionDc ?shortDescription
WHERE { ?node rdf:type dbo:Person .
            ?node foaf:name ?name. FILTER langMatches(lang(?name),'en').
            ?node dbo:abstract ?abstract. FILTER langMatches(lang(?abstract),'en').
OPTIONAL {?node dbp:shortDescription ?shortDescription. }.
OPTIONAL {?node dc:description ?descriptionDc. }.
FILTER (regex(lcase(str(?name)), \"^"+var+"\") || regex(lcase(str(?name)), \""+var+"$\") || regex(lcase(str(?name)), \" "+var+" \"))}
```

The **Ranking** and **Summarization** step is responsible for defining the most relevant elements to the domain. For this, we represent the list of propositions in the form of a graph and make use of the HARD model (Leake *et al.*, 2004), computing the weight W of each concept k, using the formula:

$$W(k) \equiv \left[\sigma \cdot TF_d(k)\right] + \left[\rho \cdot \left(\alpha \cdot a(k) + \beta \cdot h(k)\right)\right]$$

In the formula, TF_d is the frequency of the concept in the list of identified concepts in the document, *a* is the weight of the authoritative nodes (concepts with multiple incoming links), *h* is the weight of the hub nodes (concepts with multiple output links). The parameters assigned for α = 1.764 and β = 2.235 were found in the best adjustment made by Leake *et al.* (2004) and the parameters σ = 0.6 and ϱ = 0.2 were adopted in the experiment step.

[1] Stanford NER is a Java implementation of a Named Entity Recognizer. NER labels sequences of words in a text which are the names of things in the classes, such as Person, Organization and Location.

4.4 Map Visualization

The approach is characterized by the development of a proper interface for viewing the generated concept map. It includes features to display the hierarchy of concepts and to identify sub-concepts (Ausubel, 1968). Since the approach aims at automatically building concept maps from texts, most often unknown by the student, we believe that the identification of those concepts is essential for an easier understanding and assimilation of knowledge.

5 Experiments and Results

This section presents a first prototype of the approach. To perform the experiments, we use the Introduction section of the article published by Novak & Cañas (2008) as the data source. The text is written in English and is composed of 26 sentences and 617 words. Our evaluation focuses on the Element Identification step, since the approach is still being developed. Two experiments were conducted: (1) the generation of concept map containing all identified propositions extracted from the data source and (2) the generation of concept maps containing propositions filtered by the Ranking and Summarization step.

5.1 First Experiment

The first experiment identified 26 sentences, 141 propositions and 81 concepts. Figure 4, illustrates the output of this process without applying the Ranking and Summarization step, i.e., it shows all propositions extracted from the data source. A concept map constructed with CmapTools represents the process output.

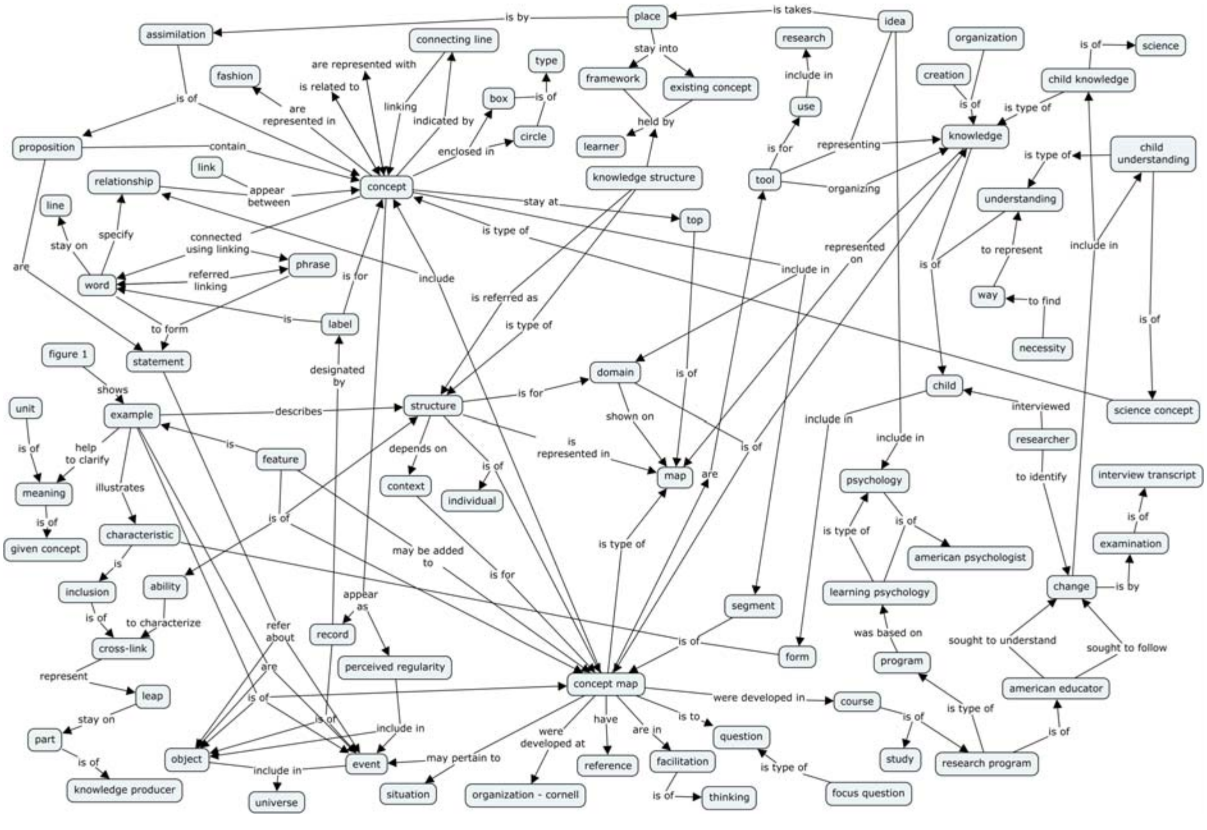


Figure 4. Extracted concept map in the first experiment.

Here, we highlight some of the features of the concept map generated by our service in the first experiment (see Figure 4 and Section 4.3):

- Proposition identification from a prepositional sentence - The proposition <*relationship*> <*appear between*> <*concept*> is extracted from the text “These are relationships or links between concepts in different segments...”. The approach creates a relationship between the concepts "relationship" and "concept" with the label "appear between". The labels are defined with the help of the prepositions mapping carried out during the Preparation step.

- Proposition identification from implicit relationships - The concept "program" is extracted from the text "This program was based on the learning psychology...", and the concept "research program" is extracted from the text "…course of Novak’s research program at Cornell...". The approach has created a specialization relationship between the concept "program" and "research program", with the label "is type of".
- Anaphora resolution - The proposition <*concept map*> <*include*> <*concept*> was extracted from the text "Concept maps are graphical tools for organizing and representing knowledge. They include concepts...". The approach associates the pronoun "they" to the concept "concept map". The approach ignores the personal pronouns of the first person, since we observe that they do not contribute significantly to the understanding of the map. Thus, the proposition <*we*> <*define*> <*concept*> extracted from the text "We define concept as a perceived regularity..." is not represented on the map. Handling demonstrative pronouns is left for future work.
- Proposition identification from distant syntactic connections - Using the syntax tree created for the text "Figure 1 shows an example of a concept map that describes the structure of concept maps and illustrates the above characteristics.", the approach extract the distant propositions: <*figure 1*> <*shows*> <*example*>, <*example*> <*describes*> <*structure*>, <*example*> <*illustrates*> <*characteristic*>, <*example*> <*is of*> <*concept map*>, <*structure*> <*is of*> <*concept map*>.
- Similarity of concepts - The concept "Ausubel", extracted from the text "The fundamental idea in Ausubel’s cognitive psychology...", and the concept "David Ausubel", extracted from "This program was based on the learning psychology of David Ausubel...", are considered as similar concepts and are represented by the most significant label, "David Ausubel". The concepts "concept" and "concepts" are associated as similar concepts and represented by the label "concept". Our approach favors the most generic or high-level labels when there are concepts with some proximity, and more specific labels otherwise. That is, the concept "good map" is represented by the more general concept "map" and the concept "interview transcript" remains with its original label.
- Labeling of entities - Concepts defined as entities of type Person are associated with their description found on DBPedia. For instance, the concept "David Ausubel" is associated with the URI "American psychologist" on DBpedia.
- Identification of multi-words concepts - The approach adopts lexical and syntactic rules to identify more complete labels of concepts, such as "knowledge producer".
- Conversion of genitive form into a normal form “is of” - The proposition <*research program*> <*is of*> <*american educator*> is extracted from "… course of Novak’s research program at Cornell...". The approach identifies and transforms the genitive form into a normal form.

5.2 Second Experiment

The second experiment added the Ranking and Summarization step to the process undertaken in the previous experiment, i.e., it applied the full process proposed by our service. The experiment used 10% of the ranking for the Summarization step, identifying 8 relevant concepts and 75 related propositions. Figure 5 illustrates the output of this process.

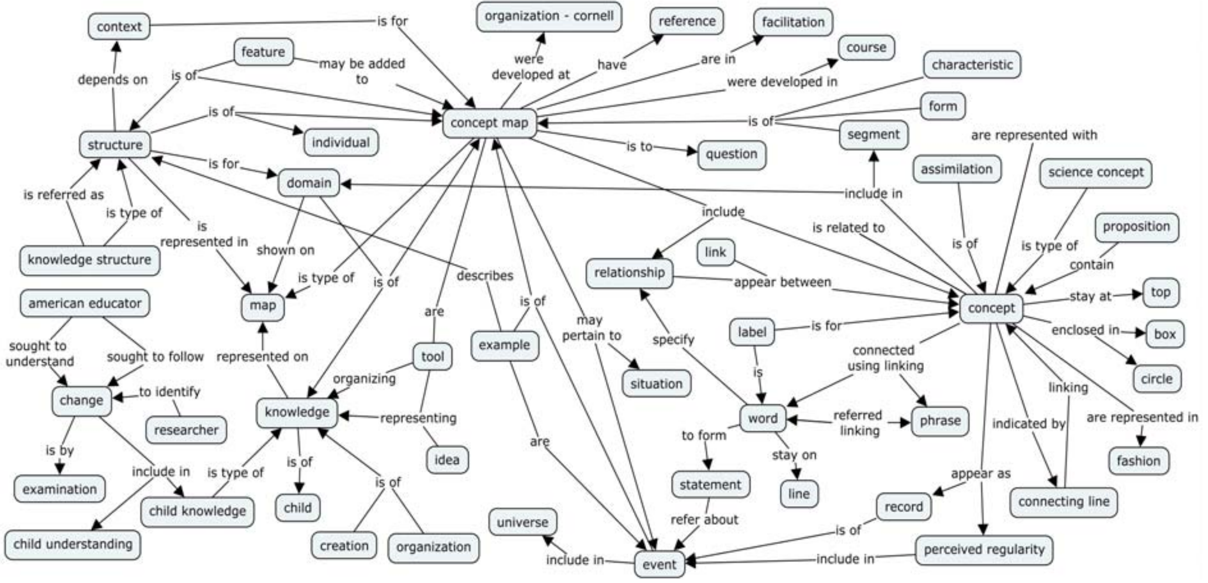


Figure 5. Concept map generated by the approach.

The experimental analysis, to date, was conducted subjectively by comparing the map built by our approach shown in Figure 5, with others from related works, applied on the same example text. Our intention is to analyze the visual quality of the map generated and the fidelity of the tool with respect to the original text.

For the visual quality, we note some strong points associated to the map built by our prototype which outperformed the results reported by related works, namely: (1) all the concepts are connected by linking phrases without fragments. Despite the Summarization step, the resulting concept map establishes relationships between concepts, even for distant concepts. (2) labels are directly extracted from the data source. (3) Neither pronouns nor named entities make up relevant concept labels. (4) Concept labels are small, meaningful, formed by multi-words expressions when applicable; (5) relationship labels are meaningful and formed by verbs; (6) relationship between concepts are sometimes not explicitly mentioned in the text; (7) concepts and propositions do not exhibit any redundancy; (8) the map faithfully represents the textual information.

In order to analyze the generated map fidelity to the text, we compared the automatically generated concept map to concept maps manually built by ten domain experts. The following instructions were provided: (1) the experts received information about the use of concept maps in general and about the purpose of the experiment; (2) they were instructed that the label of concepts and relationships should be short, meaningful and extracted from the text; (3) they were informed that concepts' labels should contain nouns, and relations' labels should contain verbs; (4) they were instructed that labels containing named entities or prepositions should be changed to more appropriate labels.

The following tables show the precision and recall calculated by comparing the map constructed by the approach with the maps generated by the experts. Table 1 shows the analysis of the identified concepts, reaching 0.47 in Precision and 0.67 in Recall. In this experiment, we disregarded the label flexion of concept maps built by experts, such as grammatical gender, number and degree.

Concept Analysis											
Expert	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5	Exp.6	Exp.7	Exp.8	Exp.9	Exp.10	**AVG**
Precision	0.60	0.42	0.44	0.56	0.52	0.48	0.46	0.38	0.42	0.42	**0.47**
Recall	0.58	0.72	0.69	0.74	0.59	0.67	0.61	0.73	0.60	0.78	**0.67**

Table 1: Results for fidelity of Concepts.

Table 2 shows the analysis of the identified relationships, obtaining 0.29 in Precision and 0.44 in Recall. In this evaluation, we consider relations as similar to those generated by the experts if they are linking the same concepts exactly and their meaning is similar. This represents a limitation of our evaluation and should be further completed by an evaluation of relations' labels in future work.

Relationship Analysis											
Expert	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5	Exp.6	Exp.7	Exp.8	Exp.9	Exp.10	**AVG**
Precision	0.37	0.24	0.23	0.33	0.31	0.35	0.23	0.23	0.29	0.28	**0.29**
Recall	0.41	0.47	0.50	0.38	0.37	0.48	0.33	0.53	0.45	0.51	**0.44**

Table 2: Results for fidelity of Relationships.

From the experiment, we note that building a map from of text is a difficult task, even for domain experts. In fact, the experts were told to not represent the concepts in their cognitive structure. Instead, they were instructed to use only the concepts expressed in the text, which they found difficult.

The results obtained in Table 1 and Table 2 are modest mainly because of the complexity of the task but also because our approach does not have an extended capability for domain identification. We believe that this issue should be alleviated by the use of a domain thesaurus. This thesaurus will be gradually constructed as new domain texts are processed. Nevertheless, we still have other challenges that can be summarized as follows:

- The anaphora resolution is still far from satisfactory, especially with respect to demonstrative pronouns, possessive pronouns and personal pronouns of the first person.
- Some assigned labels do not correspond to the labels assigned by the experts. The service, sometimes, did not make use of some adjectives and adverbs relatively important for characterizing the labels.

- Some relationships assigned by the experts were not explicitly extracted from the text because the pre-existing information in their cognitive structure interfered in their representation of the map. Thus it was not possible to directly compare them to our extracted relations.
- Some relevant domain concepts were lost during the second experiment due to our ranking.

6 Discussion and Further Work

The development of technological approaches for the automatic construction of concept maps from texts has shown promising results, although they are not yet satisfactorily resolved. In order to contribute to efforts aimed at overcoming this challenge, we present a new approach composed mainly of two elements: (1) the CMM process, which represents an extension of the version proposed by Vilallón & Calvo (2011); (2) the definition of an approach for the automatic construction of concept maps.

The approach presented in this paper stands out by the implemented grammar patterns and the use of depth-first search for the identification of concept maps' elements. Our approach relies on anaphora resolution, genitive form conversion, prepositions mapping and the identification of the type of named entities on DBpedia. For the relevance of concepts, the approach also presents a method combining the frequency of elements with the topology of the map. In general, we can state that, so far, our experiments have shown that the approach yielded acceptable results, both quantitative and qualitative, for the identification of the constituent elements of a concept map.

In our future work, we plan: (1) to enhance our anaphora resolution process; (2) to expand the classes of recognized entities and concepts based on Semantic Web knowledge bases, ensuring more appropriate concepts labeling; (3) to better identify relationships of specialization, generalization and hyponyms, allowing to establish relationships beyond the syntactic level; (4) to experiment with other ranking and scoring strategies for concept relevance. We also plan to use our prototype in a course on concept maps that will be offered to high school teachers of the public network of our state in Brazil. We count on this course to test our service in real-world settings and expect that suggestions for modifications and improvements will emerge once the course is started.

References

Aguiar, C. Z., Cury, D. & Gava, T. (2015a). Um Estudo sobre abordagens Tecnológicas para a Construção de Mapas Conceituais. Proceedings of the 20th Congresso Internacional de Informática Educativa, Chile.

Aguiar, C. Z., Cury, D. & Gava, T. (2015b). Um Abordagem Tecnológica para a Construção de Mapas Conceituais. Proceedings of the 20th Congresso Internacional de Informática Educativa. Santiago, Chile.

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1968). Educational Psychology: A Cognitive view.

Cañas, A. J., Carvalho, M., Arguedas, M., Leake, D. B., Maguitman, A., & Reichherzer, T. (2004). Mining the Web to Suggest Concepts during Concept Map Construction. In A. J. Cañas, J. D. Novak & F. M. González (Eds.), *Concept Maps: Theory, Methodology, Technology*. Proc. of the First Int. Conference on Concept Mapping, Pamplona, Spain: Universidad Pública de Navarra.

Clariana, R. B., & Koul, R. (2004). A Computer-based approach for Translating Text into Concept Map-like Representations. In A. J. Cañas, J. D. Novak & F. M. González (Eds.), *Concept Maps: Theory, Methodology, Technology*. Proc. of the First Int. Conference on Concept Mapping, Pamplona, Spain: Universidad Pública de Navarra.

Cury, D., Perin, W., & Santos Jr, P. S. (2014). CMPaaS–A Platform of Services for Construction and Handling of Concept Maps. In P. Correia, M. E. I. Malachias, A. J. Cañas & J. C. Novak (Eds), *Concept Mapping to Learn and Innovate*. Proc. of the Sixth Int. Conference on Concept Mapping, Santos, Brazil: Universidade de São Paulo.

George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol.38.

Graudina, V., & Grundspenkis, J. (2008, September). Concept Map Generation from OWL Ontologies. In A. J. Cañas, P. Reiska, M. Åhlberg & J. D. Novak (Eds.), *Concept Maps: Connecting Educators*. Proc. of the Third Int. Conference on Concept Mapping, Tallinn, Estonia: Tallinn University.

Lau, R. Y., Chung, A. Y., Song, D., & Huang, Q. (2008). Towards Fuzzy Domain Ontology based Concept Map Generation for e-learning. In Advances in Web Based Learning–ICWL 2007. Springer Berlin Heidelberg.

Leake, D., Maguitman, A., & Reichherzer, T. (2004). Understanding Knowledge Models: Modeling Assessment of Concept Importance in Concept Maps. In Proceedings of the 26th conference CSS.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In ICML, Vol. 98, pp. 296-304).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In ACL (System Demonstrations) (pp. 55-60).

Novak, J. D., & Cañas, A. J. (2008). The Theory Underlying Concept Maps and How to Construct Them. Technical Report IHMC CmapTools 2006-01, Pensacola, FL: Institute for Human and Machine Cognition. Available at: http://cmap.ihmc.us/docs/theory-of-concept-maps

Perin, W. A., Cury, D. & Menezes, C. S. (2014). NLP-Imap: Integrated Solution based on Question-Answer Model in Natural Language for an Inference Mechanism in Concepts Maps. In P. Correia, M. E. I. Malachias, A. J. Cañas & J. C. Novak (Eds), *Concept Mapping to Learn and Innovate*. Proc. of the Sixth Int. Conference on Concept Mapping, Santos, Brazil: Universidade de São Paulo.

Pinotte, G. N., Cury, D. & Zouaq, A. (2015). OntoMap: De Mapas Conceituais a Ontologias OWL. Proceedings of the 20th Congresso Internacional de Informática Educativa TISE 2015, Santiago, Chile.

Qasim, I., Jeong, J. W., Heu, J. U., & Lee, D. H. (2013). Concept Map Construction from Text Documents using Affinity Propagation. Journal of Information Science, 39(6), 719-736.

Richardson, R., & Fox, E. A. (2007). Using Concept Maps in NDLTD as a Cross-language Summarization Tool for Computing-related ETDs. In Proceedings of 10th International Symposium on ETD, Uppsala, Sweden.

Valerio, A. & Leake, D. (2006). Jump-starting Concept Map Construction with Knowledge extracted from Documents. In A. J. Cañas & J. D. Novak (Eds.), *Concept Maps: Theory, Methodology, Technology*. Proc. of the Second Int. Conference on Concept Mapping, San José, Costa Rica: Universidad de Costa Rica.

Valerio, A., Leake, D. B., & Cañas, A. J. (2008). Associating Documents to Concept Maps in Context. Title. In A. J. Cañas, P. Reiska, M. Åhlberg & J. D. Novak (Eds.), *Concept Maps: Connecting Educators*. Proc. of the Third Int. Conference on Concept Mapping, Tallinn, Estonia: Tallinn University.

Vassoler, G. A., Perin, W. A., & Cury, D. (2014). MergeMaps–A Computational Tool for Merging of Concept Maps. In P. Correia, M. E. I. Malachias, A. J. Cañas & J. C. Novak (Eds), *Concept Mapping to Learn and Innovate*. Proc. of the Sixth Int. Conference on Concept Mapping, Santos, Brazil: Universidade de São Paulo.

Villalón, J. J., & Calvo, R. A. (2011). Concept Maps as Cognitive Visualizations of Writing Assignments. Educational Technology & Society, 14(3).

Wang, W. M., Cheung, C. F., Lee, W. B., & Kwok, S. K. (2008). Mining Knowledge from Natural Language Texts using Fuzzy Associated Concept Mapping. Information Processing & Management, 44(5), 1707-1719.

Zouaq, A., Nkambou, R., & Frasson, C. (2007). Document Semantic Annotation for Intelligent Tutoring Systems: A Concept Mapping Approach. In FLAIRS Conference.