Towards Technological Approaches for Concept Maps Mining from Text

Camila Zacché de Aguiar and Davidson Cury
Federal University of Espírito Santo, UFES
Vitória, Brazil, 29075-910
{camila.zacche.aguiar | dedecury}@gmail.com

and

Amal Zouaq
University of Ottawa,
Ontario, Canada, K1N 6N5
amal.zouaq@gmail.com

Abstract

Concept maps are resources for the representation and construction of knowledge. They allow showing, through concepts and relationships, how knowledge about a subject is organized. Technological advances have boosted the development of approaches for the automatic construction of a concept map, to facilitate and provide the benefits of that resource more broadly. Due to the need to better identify and analyze the functionalities and characteristics of those approaches, we conducted a detailed study on technological approaches for automatic construction of concept maps published between 1994 and 2016 in the IEEE Xplore, ACM and Elsevier Science Direct data bases. From this study, we elaborate a categorization defined on two perspectives, Data Source and Graphic Representation, and fourteen categories. That study collected 30 relevant articles, which were applied to the proposed categorization to identify the main features and limitations of each approach. A detailed view on these approaches, their characteristics and techniques are presented enabling a quantitative analysis. In addition, the categorization has given us objective conditions to establish new specification requirements for a new technological approach aiming at concept maps mining from texts.

Keywords: **Concept Map, Concept Map Mining, Knowledge Representation.**

1 INTRODUCTION

It is well known that concept maps are resources for the representation and construction of knowledge [20], since they show, through concepts and relationships, how a subject's knowledge is organized. According to Novak [20], concepts and relationships form the basis for learning and therefore concept maps have been amply used in education in different situations and for different purposes as learning resource [12, 26, 31], means of evaluation [6, 39, 43], cross-language resource [34], representation and knowledge sharing [3, 23, 42]. In this context, concept maps are used as tools to support education, as teachers take advantage of these maps to check the student's level of understanding on a given subject, to analyze the knowledge of a classroom, to identify concepts which were not properly assimilated, or even to share the knowledge about a field of study.

The use of concept maps favors learning of the map's author who establishes systematic links between the pre-existing information and creates new and different views about them. However, the standard procedure for building a concept map involves various tasks, such as 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 concept map requires a significant time and 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 developed from the start. The

foregoing has sharpened our attention to the development of both methodological and technological bases for the automatic generation of concept maps from texts.

To facilitate the process of building concept maps, various technological approaches have been proposed to help automate that process, namely approaches for constructing concept maps from texts. These approaches require great technological and processing effort. The information extraction techniques must be able to identify concepts that are relevant to the text 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 in the text.

Thus, to better identify and analyze the features and characteristics of the technological approaches inserted in this context, this paper aims to present a categorization of the technological approaches for constructing concept maps from texts. The categorization was elaborated from a bibliographic review of the area, between the years 1994 and 2016. Subsequently, it was applied to the reviewed literature allowing for an analysis on the achieved results and elicitation of new requirements for future work.

This paper is structured into eight sections as follows: Section 2 presents the use of concept maps in learning; Section 3 discusses the representation of information using concept maps; Section 4 proposes a categorization scheme for technological approaches to constructing concept maps; Section 5 presents the study design adopted by research to investigate literature review approaches; Section 6 discusses the results obtained using this categorization to understand and compare the features of these approaches. Furthermore, it shows a more detailed study on one of the created categories, manipulation method category, due to its relatively importance in the map construction; in Section 7 a filter is applied to the categorization to identify approaches that fulfill some expected requirements to be used in our future researches on the matter; and Section 8 presents some initial conclusions of the study along with our future works.

2 ABOUT CONCEPT MAPS

Concept maps have been proposed by Novak [20] as a tool for representing and organizing knowledge, since the cognitive structure of an individual can be interpreted as a set of concepts related to each other, so as to form significant propositions. Propositions are formed as triples, consisting of two concepts connected by a link, forming a semantic unit. According to [5], a concept is defined as a regularity (or pattern) perceived in events or objects, or records of events or objects, designated by a label. Thus, a proposition is defined as a significant statement about an event or object.

On a concept map, the concepts are represented by ellipses or rectangles, and the links are represented by a directional labeled arrow. The concepts are organized hierarchically, where more generic concepts are at the highest levels, close to the root, while more specific concepts appear at lower levels, elongating to the leaves, and forming a tree-like structure. According to the theory of Meaningful Learning of Ausubel [5], human beings create meaning more efficiently when they initially consider the learning of more general and inclusive aspects, rather than working with the more specific aspects of a subject. Following this theory, the knowledge is assimilated by subsumers. A subsumer is an already stable concept, contained in the cognitive structure of an individual, which lends itself to anchor new and more specific concepts.

Fig. 1 shows the basic constituent elements of a concept map. As we examine the figure, we note that the hierarchical organization of concepts is clearly established by the position of the elements on the map. In addition, subsumers can be noted by the direction of the arrow, which may indicate the sequence and direction of the knowledge construction.

In addition to these features, the concept map is constructed based on a focal question, that is, it organizes the relevant knowledge to provide a context for the map. Although the map is built on a single focal subject, it can have different domains or segments on that particular subject. Thus, cross-links are responsible for establishing explicit relationships between concepts of different or distant subjects.

Concept maps can serve as a very useful tool for any learning theory. Thus, we can say that a concept map is a sort of non-sequential graphic representation enabling an easy understanding, construction and sharing of knowledge.

Regarding its **construction**, a concept map facilitates the transformation of tacit knowledge into explicit knowledge, since it does not require strict formats for its representation. Regarding the **understanding** of concept maps, they allow for an easy and objective way to remember pieces of information, identify relevant concepts of a domain, or view knowledge from different angles. Regarding **sharing**, maps can share knowledge representation of an area or within a group of individuals. In this case, they can be considered as an intermediate representation of a lightweight ontology. Ontology are explicit specifications of conceptualizations [52], considered *lightweight* when concepts are connected rather by general associations than strict formal connections, and *domain* when concepts formally represent the knowledge of a specific domain. In this sense, concept maps is a successful tool to elicit, assimilate and share knowledge in a particular area, be it in education or other contexts.

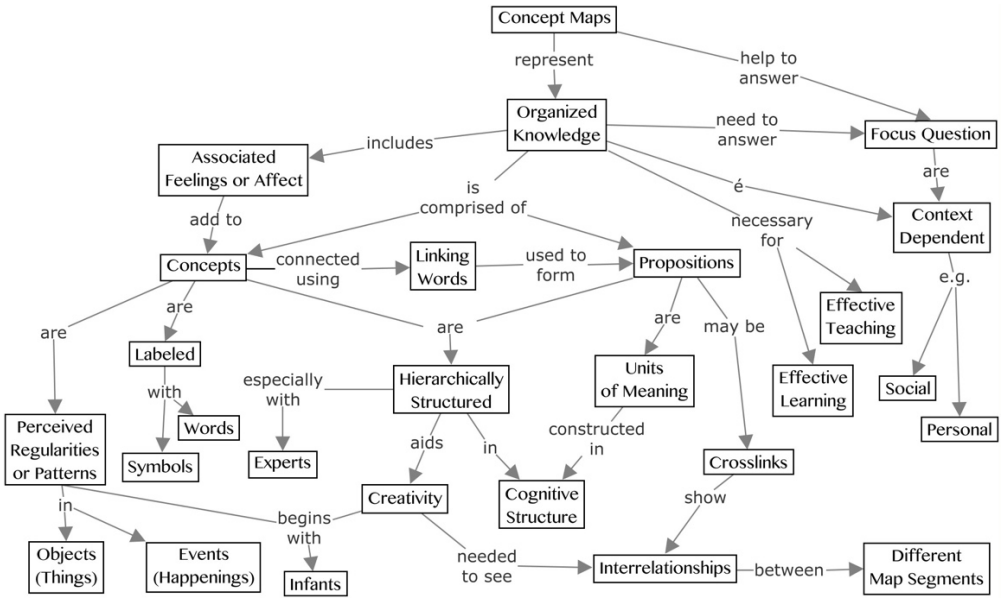


Figure 1: Example of Concept Map [5]

3 REPRESENTING INFORMATION USING CONCEPT MAPS

Information is knowledge recorded in oral, audiovisual or written form, which involves an element of meaning [48]. Therefore, information must transfer the knowledge in an orderly and adequately structured fashion. Otherwise it remains unusable and amorphous [21]. In this regard, explicit information promotes assimilation and interpretation, thus generating tacit knowledge.

One of the most used means for communicating information is the spoken or written language. Representing the information properly in written language is an arduous and expensive task. For instance, a student interested in representing tacit knowledge in a summary form would need to exert great cognitive effort to prepare the synthesis. In addition, the representation would require a sequential organization, adoption of a style, compliance with grammar rules, concern with format and others [45].

In the following, we exemplify the difference of representing information as a written text, and as a concept map (Fig. 2). In the text, the information designating a key item is represented as a concept on the map, within a box. The information that indicates an action or event is represented as a relation on the map, as a labeled directional arrow. Moreover, we note that the concept map does not represent all the information of the text, but only that forming meaningful propositions.

"Concept maps are graphical tools for organizing and representing knowledge. They include concepts, usually enclosed in circles or boxes of some type, and relationships between concepts indicated by a connecting line linking two concepts".

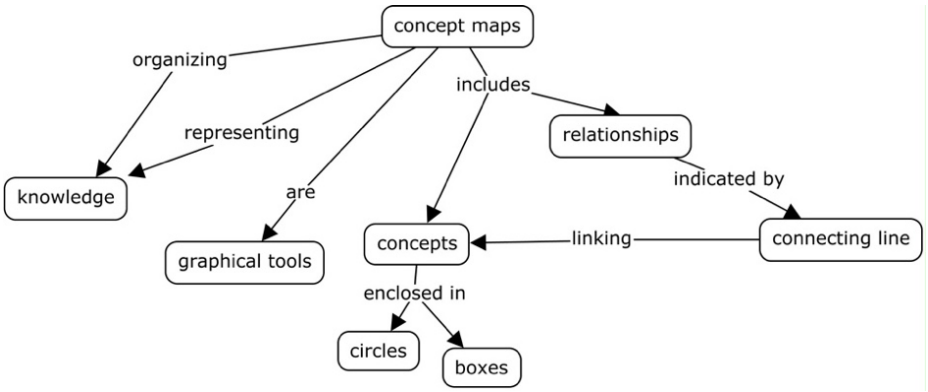


Figure 2: Written text extracted from [5] and concept map constructed from it

A written text adhering to grammatical rules can be represented by a concept map in a graphic and holistic form. In other words, a more dynamic and flexible graphical representation of the text can be constructed, which makes the assimilation and understanding of the original text easier. Furthermore, meaningful propositions allow the reader to obtain a new view point on the essential information expressed in the text. Thus, a single map can be interpreted in different ways depending on the reader, as well as, a single text can generate different maps depending on the author.

Since concept maps provide important benefits for learning, various approaches for concept map mining have been proposed [3, 22, 39, 49]. In this context, we are interested in approaches that use concept maps as a tool for the graphical representation of texts and as a learning strategy to facilitate the understanding of a complex domain. Even though concept maps can convey factual information as well as texts, the maps can be used more effectively in helping readers to build complex inferences and develop metacognitive skills [47]. In addition, the maps represent in an integrated and meaningful way, some of the most important semantics expressed in the text.

4 A Categorization of Technological Approaches for Concept Maps Mining from Text

Categorization is the process of dividing the world into groups of entities whose members are in some way like each other [51], determines the identity of concepts (categories) that are part of a domain. Therefore, this categorization is proposed with the aim to better identify and analyze the resources and characteristics of technological approaches for the construction of concept maps from texts. The categorization is defined by a model based on two perspectives and fourteen categories, which will be discussed next.

The proposed categorization is based on the perspectives identified by Aguiar & Cury [7]. They are: 1) the **Data Source**: classifies the type and quality of the input data to be used; 2) the **Graphic Representation**: establishes characteristics and rules adopted in the representation of the concept map.

The categories for each perspective, respectively, are presented in Fig. 3. These categories were identified and defined during the research, based on the bibliographic review between the years 1994 and 2016, and they are explained in this section.

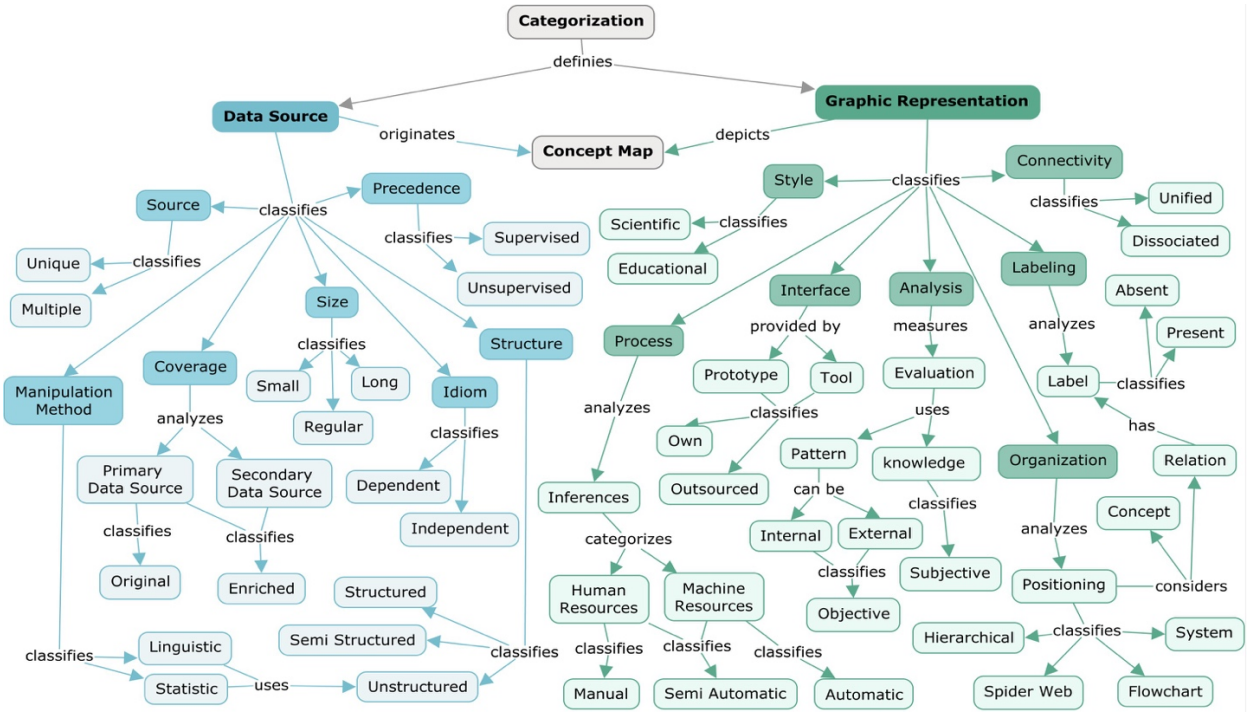


Figure 3: Concept map containing the perspectives and categories defined

4.1 Data Source

The data source is an information document used to extract the knowledge of a domain in the form of concepts and propositions. We propose categorizing the **Data Source**, restricted to written material, according to: the structure, manipulation method, idiom, size, precedence, coverage and the source, which are represented in the left area of Fig. 3.

The category **Structure** analyzes the logical structure of how information is organized in the data source. It is classified as: 1) **Structured**: shows a representation of the structure, or scheme, previously defined and homogeneous, where the data is arranged in a rigid representation and with restrictions imposed by the scheme that created them. We identified concept maps and domain ontologies as structured sources; 2) **Semi-Structured**: shows a scheme of representation defined by the document's author. It has some structure, but it is not rigid, regular, nor complete. Among the sources of semi-structured data, we consider XML [50], OWL and RDF files, since RDF and OWL are documents encoded in XML; and 3) **Unstructured**: shows no representation of structure and is generally identified as free text. It requires using natural language processing (NLP) for linguistic annotation on academic articles, theses, dissertations, queries on a domain among others.

The **Manipulation Method** summarizes the main techniques used by the reviewed approaches to extract knowledge about the data source and is strictly dependent on the type of structure. Thus, we propose two classes for the classification of the methods: 1) **Linguistic**: based on linguistic techniques [9], including, for example, linguistic pattern extraction, syntactic analysis, semantic analysis, context identification etc.; 2) **Statistic**: based on calculations of statistical measures that detect new concepts and relationships [9], including, for example, statistical analysis, co-occurrence of terms, probability, frequency, clustering etc. Some approaches offer a combination of statistical and linguistic approaches, based on syntactic parsing, linguistic filters and statistical measures.

We understand **Idiom** as the official language used for the preparation of the data source. Although idioms follow the same logical system, cultural variations have a strong influence on them and can be quite drastic in respect to the language and grammatical diversity. In such a context, the process of extracting information from the data source can be **Dependent** on, or **Independent** from the idiom used, assuming that the dependence on the idiom is closely related to the manipulation methods.

We are also interested in quantitatively analyzing the data source following the coverage, size and source. This is because these characteristics interfere with the techniques and results obtained by the approach.

The **Coverage** analyzes the origin of the data source. Most approaches adopt the **Original** coverage and consider the original data source as sufficient for the full construction of the map, from which one has a direct relation to the facts to be analyzed. Some approaches adopt the **Enriched** coverage using other secondary sources like documents retrieved from web.

The **Size** category identifies the size of the data source in terms of extension and amount of information. We can categorize the Size as: 1) **Small**: text formed by some sentences, such as an abstract; 2) **Regular**: text consisting of a few pages, such as an article, web page, didactic text and others; and 3) **Long**: text consisting of many pages, such as a dissertation and thesis.

We can classify the **Source** category as: 1) **Unique**, when the use of only one data source is necessary and sufficient for the identification and extraction of the map elements; and 2) **Multiple**, where the use of a set of data sources is necessary, either of the same structure or not. A concept map representing a document repository allows navigation in the knowledge base and exploration of the relationships between concepts. A concept map representing a unique document allows users to get a general understanding of the document.

We understand the **Precedence** as the foundation required to draw up the data source. We classify it as 1) **Supervised** when the original data source is generated or supplemented by user's knowledge. When, for example, the user needs to develop maps, annotates documents, answers questions about the domain, chooses domain ontologies, and defines list of concepts, the user's knowledge influences the definition of the data sources; and as 2) **Unsupervised** when the definition of original data source is not dependent on user's knowledge, that is, the source is the same for the expert user or not.

4.2 Graphic Representation

The construction of the concept map has a key role as a tool for the representation of knowledge. A graphic representation is more effective than a text for the communication of complex content, because the mental processing of images can be less cognitively demanding than the processing of verbal text [18]. Following this perspective, we categorize the **Graphic Representation** with respect to: analysis, process, interface, style, connectivity, organization and labeling of graphical representations of the concept map, being represented at the right area of Fig. 3.

The **Analysis** identifies the type of devices used to evaluate the results. Thus, we classify the analysis as: 1) **Subjective**: when using the knowledge of user or a domain expert to assess the outcome; 2) **Objective**: when using standards, usually statistical, as metrics to evaluating the results. This type of analysis can be replicated, given the same conditions and resulting in the same conclusion. It may be of Internal Origin, when the analysis is done with information generated from one's own source of data, or from External Origin, when the analysis is done by comparing it to other approaches.

The **Process** analyzes the type of interventions that occur throughout the construction of the concept map and can be classified as: 1) **Automatic**: when the intervention occurs only with machine resources from the choice of the data source to the construction of the concept map; 2) **Semi-Automatic**: a mixture of human and machine intervention.

Thus, the automatic intervention is used to generate propositions and human intervention to construct the map, or vice-versa; as 3) **Manual**: the human intervention is critical throughout the process, although some activities are performed by automatic intervention, as seen in approaches that generate candidate concepts automatically, but leave to the user the construction of propositions and the graphical representations.

The **Interface** makes explicit the relative position of each concept within the map. Given the importance of graphical view, we believe that any approach needs an interface, either its **Own** or **Outsourced**, when using resources which do not belong to the approach. In this last case, it adopts consolidated tools like CmapTools [9], Graphviz [22] and WebDot [30].

The **Labeling** analyzes the presence of the linking words or labels that specify the relationship between the concepts of the proposition. We can classify them as: 1) **Present**: when there is the presence of labels on the relations. It can be subdivided into Open label, when the label is extracted from all possible relationships in the text, such as sentence predicate; and Closed label, when the label is extracted from a closed set of relations, such as stereotype; and 2) **Absent**: when there is no presence of labels.

The **Connectivity** analyzes the ability to establish links and cross-links in the construction of the concept map. In this context, we classify connectivity as: 1) **Unified**: establishes cross-links relations between the subdomains of knowledge represented on the map, showing how they relate to each other in a single interconnected map. In other words, there is no portion of the map unplugged from the map as a whole; and as 2) **Disassociated**: establishes no cross-link relation in order to represent various portions of maps not connected. These are observed in approaches that fail to uncover the link between some concepts or that cannot create the links.

The **Style** determines the type of the concept map to be built. We classify the style as: 1) **Educational**: when such rules are irrelevant. Usually maps of this kind are developed by children in order to represent what they know about something; and 2) **Scientific**: built from a data source resulting from any scientific research. It is governed by two basic rules: the map might contain only concepts, and there is always a verb in a relationship between concepts. In this case maps are used for the development of ontologies, interoperability, organizational memory etc.. A concept map of scientific style is directed to a specific purpose, such as evaluation and support for learning, representation and summarization of the text among others.

Following are some examples to illustrate the category style. A child writes the sentence "*Mary is beautiful*". The sentence can be represented by a simple concept map containing the triple (*Mary, is, beautiful*). Nevertheless, we know that neither "*Mary*" nor "*beautiful*" are concepts. *Mary* can be defined as instance of person or woman, and *beautiful* as a property of *Mary*. However, the sentence represents the knowledge constructed by a child and it is important to be represented in a concept map of an educational style. This is also the case of "*a bee can fly*", "*John loves Mary*" and many others. Consider the following sentence now: "*Teachers teach certain subjects*". A concept map containing the triple (*teachers, teach, certain subjects*) represents more clearly the significant relationship between undoubtedly two concepts. In this case, the map stems for the scientific style.

Based on Tavares [38] we analyzed the **Organization** of the elements on the map generated by approaches according to: 1) **Hierarchical**: identified in most approaches, it organizes the concepts in order of importance, locating the more general at the top of the map; and 2) **Spider web**: organizing the central and most important concept in the middle of the map; 3) **Flowchart**: not identified in any of the studied approaches, organizes the concepts linearly including start and end points; and 4) **System**: not identified in any of the studied approaches, organizes the concepts as a flowchart, and adds input and output concepts. Some approaches may take more than one type of organization, as noted in [30], whose map organization, hierarchical or spider web depends on the purpose of the author.

5 STUDY DESIGN

A review of literature was conducted to map the studies that address technological approaches for the construction of concept maps from texts. Since the state of the art of concept map mining does not comply with any standard guidelines it is difficult to categorize related issues. This study aims at providing a more systematic analysis scheme of the works in this context.

This study was conducted following the guidelines suggested by Petersen *et al.* [24]. The study consists of the following steps described respectively in sections 5.1, 5.2 and 6: 1) defining research questions, 2) conducting research on primary studies, 3) data extraction, and 4) data analysis.

5.1 Research Questions

The initial question that motivated this categorization was: *Which technological approaches are being developed for the construction of concept maps from texts?* The following research questions were defined:

- *(RQ1)* What are the main characteristics of technological approaches in this context?

- *(RQ2)* What are the main characteristics of the concept maps built by these approaches?
- *(RQ3)* What is currently known about the benefits, challenges and limitations of the approaches?
- *(RQ4)* Which methods and techniques exist to support the development of these approaches?
- *(RQ5)* What evaluations should be designed to assess the concept maps built by these approaches?

5.2 Research on the Primary Studies

Starting from these research questions, we defined search sources, as well as inclusion and exclusion criteria. The search strategy included only electronic databases, and they are: IEEEXplore Digital Library, ACM Digital Library, and Elsevier Science Direct. On these search sources, the following keywords were used:

> ("concept map" OR "concept mapping" OR "concept maps" OR “concept map mining”) AND ("construction" OR "constructing" OR "creation" OR “creating” OR "generation" OR “generating” OR “building”) AND ("automatic" OR "automated" OR "automatically")

Initially the selection of potentially relevant studies was determined by the analysis of the title, keywords and abstract. After that, the selection of the studies was determined by reading the whole paper.

For the inclusion of the study, the following criteria were considered:

- *(IC1)* The work´s different versions published by an author on the same approach.
- *(IC2)* Studies written in English or Portuguese language.
- *(IC3)* Studies that address some of the research questions.

For the exclusion of the study, the following criteria were considered:

- *(EC1)* Repeated studies. If a study is available in more than one search source, it will be considered only the first time it is found.
- *(EC2)* Non-scientific studies (notes, index, editorials, prefaces).
- *(EC3)* Irrelevant studies for the research.
- *(EC4)* Studies whose files could not be accessed by the institution.

After applying search string to search sources, 134 articles were returned. After downloading, only 55 papers were considered potentially relevant in the first selection. In the second selection, a better analysis on the primary studies was conducted, where all papers were read and 30 relevant papers were selected. Table 1 summarizes the selection process and presents the number of papers identified at each step.

Table 1: Selection process of primary study

Source	Studies Retrieved	1º Selection	2º Selection				
			Irrelevant	Repeated	Non-Scientific	Non-Access	Primary Study
IEEE Xplore	*94*	*33*	*19*	*4*	*0*	*0*	*10*
ACM	*19*	*11*	*6*	*0*	*1*	*0*	*4*
Science Direct	*21*	*16*	*2*	*1*	*0*	*0*	*14*
Total	***134***	***55***	***27***	***5***	***1***	***0***	***30***

Although the search was not limited to a particular period, all studies were found between the years 2001 and 2016. The graph of Fig. 4 illustrates the concentration of studies per year. We can observe that the highest concentration of studies of this area occurred in the years 2008, 2009 and 2012.

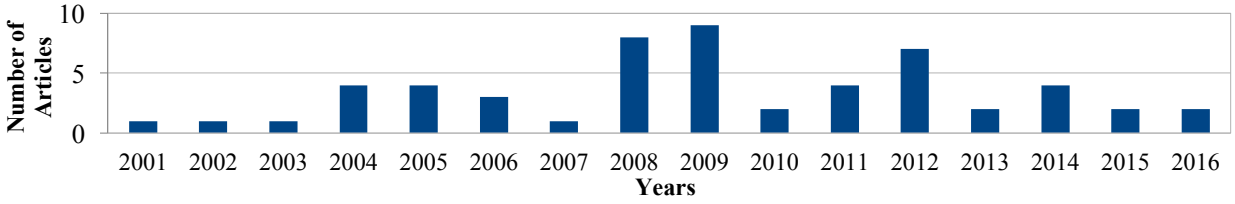


Figure 4: Concentration of studies per year

6 ANALYSES OF THE RESULTS

To answer the research questions in Section 5.1, the categorization proposed in Section 3 was adopted as a metric for analyzing the technological approaches selected in the primary study. Table 2 synthesizes the result of the categorization performed for the 30 selected papers, divided into two main areas: 1) *Categories*: located on the left side, horizontally arranged, associates the Reference area with the categories of Data Source, and Graphic Representation; 2) *References*: on the right side, arranged vertically, denotes the approach identified by its number in the list of references at the end of this article.

To understand the data represented in Table 2, it is necessary to know that each reference is classified individually for each category and the data analysis should be performed crosswise. Therefore, for each reference located vertically, there is a category located horizontally that is directly associated. To represent that the reference satisfies the category located in the left area the notation "■" is adopted and to represent that the reference does not satisfy this category, an empty space is adopted.

Table 2: Categorization applied to the approaches of primary study

Categories			References																														
			49	*39*	*3*	*17*	*43*	*34*	*4*	*26*	*12*	*22*	*27*	*8*	*16*	*19*	*31*	*40*	*41*	*6*	*42*	*33*	*1*	*25*	*32*	*28*	*35*	*44*	*30*	*23*	*13*	*2*
DATA SOURCE	**Precedence**	*Supervised*		■	■				■	■	■						■	■	■	■	■		■	■	■	■		■	■	■		
		Unsupervised	■			■	■	■				■	■	■	■	■						■					■				■	■
	Idiom	*Dependent*	■			■		■	■	■		■		■	■							■				■	■			■	■	■
		Independent		■	■		■				■		■				■	■		■	■		■		■			■				
	Structure	*Structured*																														
		Unstructured	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
		SemiStructured																														
	Amplitude	*Natural*	■	■	■	■		■	■			■	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■
		Enriched					■			■	■															■						
	Source	*Unique*	■			■	■	■	■	■					■				■			■				■				■	■	
		Multiple		■	■						■	■	■	■		■	■	■		■	■		■	■	■		■	■	■			■
	Size	*Small*		■						■			■	■	■			■		■	■		■	■	■		■	■				
		Regular	■		■	■	■				■	■				■	■		■			■				■			■	■	■	■
		Long						■	■																							
	Manipulation Method	*Linguistic*	■		■	■		■	■	■		■			■							■				■	■			■	■	■
		Statistic	■	■	■	■	■	■			■	■	■	■		■	■	■	■	■	■		■	■	■		■	■				
CONSTRUCTION	**Connectivity**	*Unified*	■	■			■	■		■						■		■		■	■	■	■		■	■	■	■	■	■	■	■	■
		Dissociated			■									■			■		■					■					■				
	Style	*Educational*																														
		Scientific	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	Purpose	*Learning Evaluation*		■			■						■					■		■	■	■		■	■		■	■				
		Text Summarization						■											■											■	■		
		Learning Support								■	■					■	■					■										
		Text Representation	■		■	■			■			■		■	■								■			■			■			■
	Organization	*Hierarchical*	■			■	■	■									■			■	■	■		■	■			■	■	■		
		Graph		■											■	■		■	■				■			■					■	
		Spider Web								■																	■		■			■
	Analysis	*Objective*	■																■													■
		Subjective			■	■	■	■	■	■	■	■		■	■	■	■		■	■		■			■	■	■	■	■	■	■	■
	Interface	*Own*	■	■	■	■	■	■		■	■			■		■			■	■		■	■	■		■	■	■			■	■
		Outsourced										■					■												■	■		
	Labeling	*Present*	■			■				■																				■		■
		Absent		■	■		■	■						■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	
	Process	*Manual*																														
		SemiAutomatic									■	■				■																
		Automatic	■	■	■	■	■	■	■	■			■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

From the perspective **Data Source,** we can observe in the category **Structure** that most approaches adopt *Unstructured* (100%) sources, since text is the focus of this study. In the category **Precedence** we found that some approaches choose the *Supervised* (56%) to better extract the author's contributions for the identification of the map elements.

In the category **Coverage** we concluded that most of the approaches use *Original* (86%), that is, extract the elements directly from their data source. Nevertheless, some approaches have sought the web for new knowledge to enrich the map. However, the difficulty of finding and extracting relevant information within the vast web restricts many approaches. Looking at the category **Source**, we identify the source *Multiple* (63%) as the most used, in this case, the approaches are interested to represent the knowledge of a domain, or a group of individuals, about a domain. Looking at the category **Size**, we found that most approaches use a *Regular* (50%) size text, because the approach neither need to have high processing power as long texts nor greater precision as small texts.

According to the **Manipulation Method**, we note that most approaches adopt *Statistical* methods (50%), some adopt *Linguistic* methods (30%), and only a small portion adopts both methods (20%). As the category **Idiom** is strictly dependent on the manipulation method used, some approaches are *Dependent* on the idiom (46%) such as English (85%), Spanish (7%), and Croatian (14%).

From the perspective **Graphic Representation**, we observe that many approaches assume some characteristics of maps in Novakian style, adopting together **Connectivity** as *Unified* (63%), where all the propositions are connected and do not have fragments of the map; and **Organization** as *Hierarchical* (43%), positioning concepts with a certain hierarchy on the map. However, the approaches do not adopt **Labeling** as *Present* (16%), where there is the presence of labels on the relationships, using mostly Labeling as *Absent* (70%), that is, without the presence of labels.

According to the category **Process**, we identify that the *Automatic* (90%) is the most used by the approaches. Although it does not show the best result, this process is user independent. From the analysis in the category **Interface**, we observe that most approaches develop their *Own Interface* (66%). Due to difficulties in analyzing a technological approach for the construction of concept maps, the majority adopts a *Subjective* **Analysis** (73%), delegating the responsibility assessment to an expert. Although it is the most widely used, it is not the most appropriate, because it makes it impossible to validate or replicate the analysis.

In the category **Style**, we can observe that 100% of the approaches studied are of *Scientific Style*, since maps containing a known guideline are better suited for comparative studies, evaluation and learning. In addition, we have observed approaches that aim at student's evaluation (36%), graphical representation of text (36%), learning support (16%) and summarization of text (13%).

Based on the categorization and analysis presented, we can observe some advantages and disadvantages of the studied approaches. We have identified some characteristics signaled by the categories:

- **Precedence**: Most approaches adopt the category Supervised and hence they limit the construction of the map to a previously known domain.
- **Purpose**: Approaches to constructing concept maps from texts have been developed with the purpose of evaluating learning and representing text.
- **Source**: Approaches adopt multiple data sources, that is, more than a text, since it is more accurate to identify relevant concepts from a set of data sources.
- **Interface**: Although most of the approaches adopt their own interface, they do not develop the interface potential for learning beyond the graphic representation.
- **Labelling**: although the identification of relation labels is relevant to the construction of a map, many approaches still define absent labels. In this case, the map does not represent the meaningful propositions; instead it represents the relation's force between relevant concepts in the text.
- **Connectivity**: although most approaches build unified maps, ensuring this feature is a challenge that in most approaches is related to the text.

6.1 On the Evaluation

The evaluation proposed by the various approaches to assess the generated concept maps (CMG) can be either objective or subjective. The graph of Fig. 5 illustrates the types of assessments observed in our primary studies, where the objective, subjective and non-evaluation are represented in green, blue and orange color, respectively.

Analyzing the graph, we can observe that most approaches do not use an objective evaluation (7%). Furthermore, they generally do not perform an assessment of the quality or accuracy of the concept map (40%). Among the

approaches studied, only one carried out an objective analysis comparing the propositions extracted by the approach with the annotated propositions in a corpus.

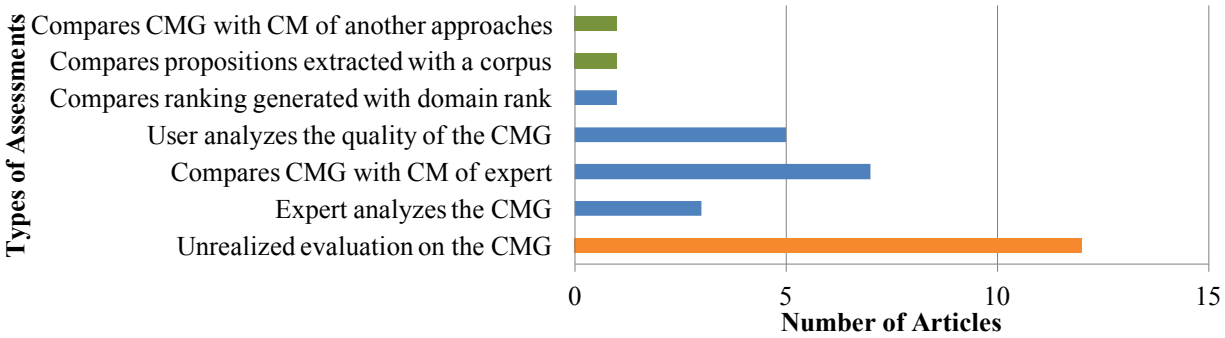


Figure 5: Type of assessments performed by the studies

6.2 On the Manipulation Methods

We can consider that the Manipulation Method category strongly influences the outcome of the approach, being totally dependent on the techniques applied in the data source for the extraction of knowledge. Thus, we can synthesize the information extraction process to build a concept map in four steps:

The **Pre-Processing** step changes the data source to allow the mining process to extract more intelligible information such as removing formatting, removing special characters and eliminating label markers, tags and font style.

The **Normalization** step proposes a semantic approximation of terms, in order to reduce the ambiguity and term variation. This comprises:

- Stemming or lemmatization. Lemmatization is used to find the "lemma" of the word, disregarding grammatical changes such as tense and plurality [10]. The main purpose of stemming is to reduce different grammatical forms to the "root" form.
- Co-reference resolution, it is the task of finding all expressions that refer to the same entity in a discourse [14]. It is common that the candidates compete to be the antecedent of an anaphor [36].
- Named entity recognition. A named entity is a sequence of words that designates some real-world entity, such as "Brazil," "UFES" and "Steve Jobs". Named entity recognition identifies mentions in text belonging to predefined types, such as person, organization and location [11].
- Stop words deletion as well as the removal of all information that does not constitute knowledge in the text;
- Multi-words and acronym identification;
- Synonymy and related concept detection using a dictionary.

The **Elements Identification** step selects candidate terms for concepts and relationships in order to form future propositions on the map. Statistic-based approaches handle documents by means of metrics and numbers, however they may suffer unpredictable results and semantic loss. The purely linguistic-based approaches are more accurate than the statistical ones though, in most cases, they are based on external knowledge databases. For these purposes, different techniques are adopted for each type of approach.

For linguistic approaches, we can point out the use of patterns and rules on the grammatical structure of text, such as:

- Tokenization is the process of converting a sequence of characters (text) into a sequence of meaningful units (words) that compose the text. The term token is used to designate these units, which correspond to one or more textual expressions such as "27/01/2017", "100,00" and "pre-processing".
- Morphological Analysis is focused on the individual terms. For each word in a sentence the analysis identifies its grammatical class, morphological class or part of speech (noun, verb, preposition etc.) and its flexion (gender, number and grade).
- Syntactic analysis is focused on the relationship between words according to a certain grammar theory. The analysis produces a full parse tree from a sentence. From the parse, we can find the relation of each word to

all the others in the sentence, and typically also its function in the sentence. The syntactic analysis may be divided between the constituency and dependency grammars [37].

For statistical approaches, we can point out the use of clustering and statistical techniques to identify terms for the domain:

- Clustering is a descriptive task in which one seeks to identify a finite set of clusters to describe the data [29] based on associating among features within the data, on the contexts they have in common [46]. Usually used to discover group of relevant concepts.
- The frequency of terms assumes that the weight or relevance of a term occurring in a document is proportional to its frequency [15].
- Association rules are created by analysing data for frequent if/then patterns. It uses the criteria of how frequently the items appear and number of times the if/then statements were found.

The **Summarization** step is responsible for reducing the identified elements, defining the most relevant ones for the data source. Usually the approaches adopt a domain ontology, frequency in the text, or ranking algorithm to identify the most relevant concepts.

From an analysis of the approaches of the primary study, we identified a set of techniques used for the extraction of information. Table 3 synthesizes the main techniques identified.

The table is divided into two main areas: 1) *Techniques*: located on the left side, horizontally arranged, associates the Reference area with the techniques identified; 2) *References*: on the right side, arranged vertically, denotes the approach identified by its number in the list of references at the end of this article.

Table 3: Techniques identified in the Approaches

Techniques	References																												
	49	*39*	*3*	*17*	*43*	*34*	*4*	*26*	*12*	*22*	*27*	*8*	*16*	*19*	*31*	*40*	*41*	*6*	*42*	*33*	*1*	*25*	*32*	*28*	*35*	*44*	*30*	*23*	*13*	*2*
Pre-Processing	■	■										■	■	■	■									■			■	■	■	■
Normalization																														
Stopword										■		■		■						■					■			■	■	■
Stemming										■				■	■					■					■		■		■	
Lemmatization										■		■																■		
Acronym										■			■		■												■			
Synonymous	■														■									■	■		■	■	■	
Anaphora Resolution	■			■																				■				■	■	
Entity Recognition										■																		■		
Similarity of Terms					■							■	■																	
Element Identification																														
Tokenization	■											■	■							■				■	■			■	■	■
Lexical Analysis	■			■		■	■			■			■							■				■	■			■	■	■
Syntactic Analysis	■			■		■	■			■			■							■				■				■	■	■
Syntactic Dependency	■						■			■			■							■				■				■	■	
Semantic Dependency				■																				■						■
Grammar Pattern																													■	■
Association Rules		■														■	■	■		■	■	■				■				
Terminology Map																														■
Graph Theory																														■
Neural Network			■																											
Clustering			■	■					■	■		■			■			■									■			
Fuzzy Taxonomy		■									■								■		■		■		■					
Frequency of Terms	■		■		■	■			■	■	■	■		■	■	■		■			■		■		■	■	■	■		
Frequency of Link					■				■																					
Co-occurrence of Terms												■		■		■	■		■			■								
Burst of Word																	■													
Proximity Position										■					■					■				■	■		■			

Ranking Algorithm																		■	
Thesaurus																		■	
Ontology		■												■					
Knowledge Database		■			■														

Noting the analysis presented in the table, we highlight the following features:

- Approaches adopting a linguistic method correspond to only 36% of the studied approaches, since this method requires more computational effort and have some limitations.
- Normalization techniques are used both by linguistic and statistical methods.
- Terms frequency techniques (60%) are primarily used by statistical methods, however they can be adopted in linguistic methods for the identification of relevant elements.
- Some approaches use knowledge base (7%), ontologies (7%) or thesaurus (3%) as source for the identification of elements belonging to a domain.

7 APPROACHES IDENTIFIED FROM THE CATEGORIZATION

By adopting the proposed categorization scheme, it was possible to analyze the different approaches with better defined and objective metrics. This allowed for a better understanding and comparison of the characteristics presented by each approach. In addition, we are interested in using the categorization to objectively identify the set of approaches that fulfill certain requirements. Therefore, the following are the criteria used for the selection of approaches directed to the construction of maps from texts that fulfill our future research interests:

- Approach that uses only a single unstructured text of regular size.
- Approach that does not use the knowledge of the user to modify the data source and accepts text from any domain;
- Approach that adopts only machine resources;
- Approach that generates maps containing labels for concepts and relationships. Moreover, the resulting map must not have fragmented portions;
- Approach that represents only the information contained in its own data source.

To fulfill these requirements, the following filters were applied to the categorization conducted on the literature review in Section 6:

> Style (*Scientific*), Purpose (*Summarization*), Precedence (*Unsupervised*), Idiom (*all*), Structure (*Unstructured*), Coverage (*Original*), Source (*Unique*), Size (*regular or long*), Manipulation Method (*all*), Connectivity (*Unified*), Organization (*all*), Analysis (*all*), Interface (*all*), Labeling (*Present*) and Process (*Semi* or *Automatic*).

Since none of the approaches fulfilled the requirements of the mentioned filters, we kept only Style, Structure and Labeling. Accordingly, four approaches fulfilled these requirements and are synthesized in Table 4 where the right side provides an overview of the approaches and the left side shows the concept map constructed by them.

Looking at the map generated by the approaches in the Table, we observe: map is fragmented in portions [49]; the approach assigns very long labels to concepts [49]; it accepts pronouns as labels for concepts [49]; it accepts prepositions as labels for relationships [49]; it accepts relationships without label [23]; map is created on a specific domain [26, 23]; the map is closer to representing the domain of the subject than the real content of the text [26]; map is created using a set of domain documents [2] or a small text containing some sentences [26]; it uses other domain data source in addition to the text such as ontology [2, 26], knowledge base [26] and thesaurus [23]; and does not show the direction (uses no arrow) of the association between concepts [26].

According to the analysis carried out on these approaches, we can identify the following challenges in: 1) defining small and meaningful labels; 2) identifying relevant domain concepts; 3) establishing links between concepts which are not evidenced in the text; and 4) identifying the domain of a document. Such situations are still challenging for the automatic generation of concept maps.

Table 4: Approaches identified from the Categorization

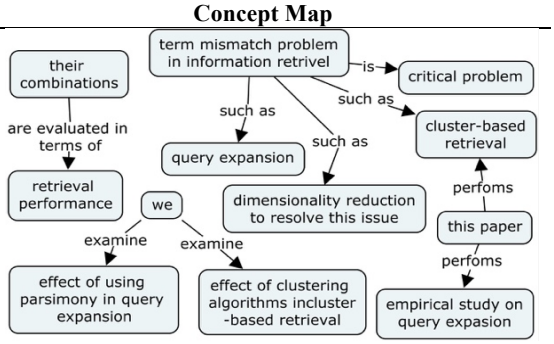
Approaches Detail	Concept Map
The approach [49] generates concept maps from abstracts in English. This approach uses morphological and syntactic analysis, identifying the elements based on the structure of the phrases and syntactic rules. It applies normalization to correct orthographic mistakes, and 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.	

Figure 6: Concept map generated by [49]

The approach [23] generates maps from legal documents in Croatian language as a summarization of the text. This approach creates hierarchical maps from a specific area using domain thesaurus. From a domain corpus, the documents are preprocessed and the metadata is mapped. It uses linguistic techniques for lemmatization, entity recognition, co-reference resolution, lexical and syntactic analysis. Concepts are identified by the metadata and frequency of terms in the text. Propositions are extracted from the subject-predicate-object pattern in the sentence containing the identified concepts and by the relationships established between the concepts in the thesaurus. A tree structure formed by 25-30 concepts of the propositions is constructed hierarchically assigning the text title as root node.

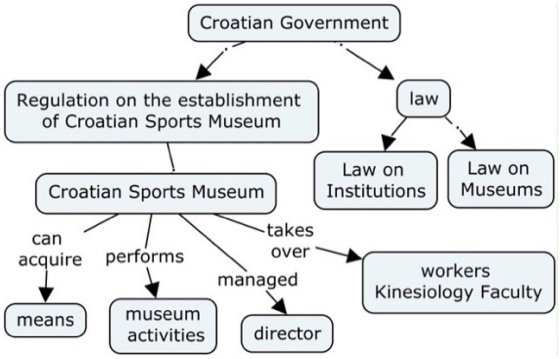


Figure 7: Concept map generated by [23] (translated)

The approach [2] generates concept maps from texts in English language as an intermediate step for generating an ontology. For that, it uses linguistic techniques for segmentation, normalization with stemming and syntactical analysis. It applies machine learning to identify keywords and creates a semantic concept map of sentences containing these keywords. The triples are extracted from syntactical rules and grammatical dependencies between the words in the sentence. Lexical-semantic patterns interpret this structure to extract concepts and relationships. Finally. it performs statistical analysis to define the relevance of concepts and relations.

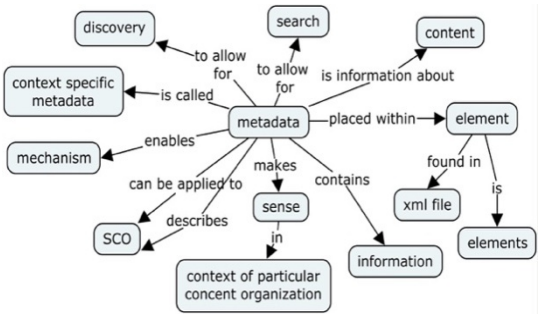


Figure 8: Concept map generated by [2] (clipping)

The approach [26] generates concept maps from clinical text in English language. This approach uses concepts and an ontology to obtain rich information about the domain. The system pre-processes a set of medical terms compiled into lists and search for domain terms in text. The user chooses a concept and queries to retrieve information about the concept in the knowledge bases using lexical and semantic resources.

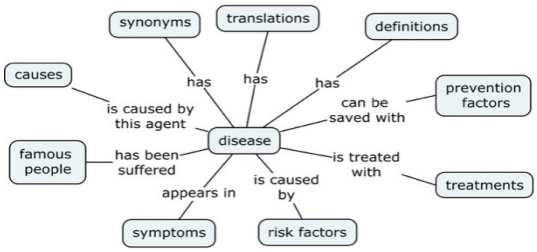


Figure 9: Concept map generated by [26]

8 SOME PRELIMINARY CONCLUSIONS

We consider categorization as a model or method to organizing the objects of a given universe into groups, for a specific purpose. In this work, we use the categorization as a means of comparing and evaluating the effectiveness and efficiency of technological approaches for the automatic construction of concept maps. After continuous reading

of several studies found in three databases (IEEE Xplore, ACM and Elsevier Science Direct) between 1994 and 2016, we observed that each study adopted particular criteria to describe the proposed approach. The initial question that motivated all this research unfolded in five other questions and all of them were positively answered throughout the work. The entire study was guided by Petersen's guidelines, explained in detail in Section 5.

We believe that our categorization is consistent and has added important information to the understanding, development and analysis of the approaches for concept map mining from text. However, it is still incomplete. Due to the amplitude and extent of the category Manipulation Method, we limited the classification of the approaches to the type of method adopted, although we have inspected some of the techniques used.

We aim that the use of an objective categorization to understand and compare approaches in this context is a first step to improve and expand research in the area. The lack of metric becomes impractical to compare and evaluate the effectiveness and coverage of proposals. Therefore, looking at the benefits that the automatic construction of concept maps brings to the learning, we intend to encourage further investigations from this contribution. As of now, our categorization is already supporting our next step, which consists of the elaboration of a conceptual model for a new technological approach aiming at the automatic summarization of concept maps.

References

[1] A. Aajli, & K. Afdel. *A new hybrid approach for constructing the concept map based on fuzzy prerequisite relationships*. In 2014 Third IEEE International Colloquium in Information Science and Technology (CIST) (pp. 115-121). IEEE. 2014. DOI: 10.1109/CIST.2014.7016604.

[2] A. Zouaq, & R. Nkambou. *Evaluating the generation of domain ontologies in the knowledge puzzle project*. IEEE Transactions on Knowledge and Data Engineering, 21(11), 1559-1572. 2009. DOI: 10.1109/TKDE.2009.25.

[3] A. Pipitone, V. Cannella, & R. Pirrone. *Automatic concept maps generation in support of educational processes.* Journal of e-Learning and Knowledge Society, 10(1), 2014.

[4] A. M. Olney, W. L. Cade, & C. Williams. *Generating concept map exercises from textbooks*. In Proceedings of the 6th workshop on innovative use of NLP for building educational applications (pp. 111-119). Association for Computational Linguistics, 2011.

[5] A. Cañas, M. J. Carnot, P. Feltovich, R. R. Hoffman, J. Feltovich, & J. D. Novak. *A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support*. Pensacola, FL. 2003.

[6] C. H. Lee, G. G. Lee, & Y. Leu. *Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning.* Expert Systems with Applications, 36(2), 1675-1684. 2009. DOI: 10.1016/j.eswa.2007.11.049.

[7] C. Z. Aguiar, & D. Cury. *A categorization of technological approaches to concept maps construction*. In Learning Objects and Technology (LACLO), Latin American Conference on Learning Objects (pp. 1-9). IEEE. 2016. DOI: 10.1109/LACLO.2016.7751743.

[8] C. Lipizzi, D. G. Dessavre, L. Iandoli, & J. E. R. Marquez. *Towards computational discourse analysis: A methodology for mining Twitter backchanneling conversations*. Computers in Human Behavior, 782-792, 2016. DOI: 10.1016/j.chb.2016.07.030.

[9] C. C. C. Pérez, & R. Vieira. *Mapas Conceituais: geração e avaliação*. In Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana. 2005.

[10] D. Biber, S. Conrad, & R. Reppen. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press. 1998.

[11] D. Nadeau, & S. Sekine. *A survey of named entity recognition and classification*. Lingvisticae Investigationes, 30(1), 3-26. 2007. DOI: 10.1075/li.30.1.03nad.

[12] D. B. Leake, A. Maguitman, T. Reichherzer, A. J. Cañas, M. Carvalho, M. Arguedas and T. Eskridge. *Aiding knowledge capture by searching for extensions of knowledge models*. In Proceedings of the 2nd international conference on Knowledge capture (pp. 44-53). ACM, 2003. DOI: 10.1145/945645.945655

[13] E. L. Karannagoda, H. M. T. C. Herath, K. N. J. Fernando, M. W. I. D. Karunarathne, N. H. N. D. De Silva, & A. S. Perera. *Document analysis based automatic concept map generation for enterprises.* In Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on (pp. 154-159). IEEE. 2013. DOI: 10.1109/icter.2013.6761171.

[14] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, & D. Jurafsky. *Deterministic coreference resolution based on entity-centric, precision-ranked rules*. Computational Linguistics, 39(4), 885-916. 2013. DOI: 10.1162/coli_a_00152.

[15] H. P. Luhn. *A statistical approach to mechanized encoding and searching of literary information*. IBM Journal of research and development, 1(4), 309-317. 1957. DOI: 10.1147/rd.14.0309

[16] I. Bichindaritz, & S. Akkineni. *Concept mining for indexing medical literature.* Engineering Applications of Artificial Intelligence, 19(4), 411-417, 2006. DOI: 10.1007/11510888_68.

[17] I. Qasim, J. W. Jeong, J. U. Heu, & D. H. Lee. *Concept map construction from text documents using affinity propagation*. Journal of Information Science, 39(6), 719-736. 2013. DOI: 10.1177/0165551513494645.

[18] I. Vekiri. *What is the value of graphical displays in learning?*. Educational Psychology Review, 14(3). 2002.

[19] J. H. Lee, & A. Segev. Knowledge maps for e-learning. Computers & Education, 59(2), 353-364, 2012. DOI: 10.1016/j.compedu.2012.01.017

[20] J. D. Novak, & A. J. Cañas. *A teoria subjacente aos mapas conceituais e como elaborá-los e usá-los*. Práxis Educativa, 5(1), 9-29. 2010. DOI: 10.5212/praxeduc.v.5i1.009029.

[21] K. McGarry and H. V. de Lemos. *O contexto dinânico da informação: uma análise introdutória*. Briquet de Lemos, 1999.

[22] K. Zubrinic, D. Kalpic, & M. Milicevic (2012). *The automatic creation of concept maps from documents written using morphologically rich languages*. Expert systems with applications, 39(16), 12709-12718. DOI: 10.1016/j.eswa.2012.04.065

[23] K. Žubrinić, I. Obradović, & T. Sjekavica. *Implementation of method for generating concept map from unstructured text in the Croatian language.* In Software, Telecommunications and Computer Networks (SoftCOM), 2015 23rd International Conference on (pp. 220-223). IEEE. 2015. DOI: 10.1109/softcom.2015.7314098.

[24] K. Petersen, R. Feldt, S. Mujtaba, & M. Mattsson. *Systematic mapping studies in software engineering*. In 12th international conference on evaluation and assessment in software engineering (Vol. 17, No. 1). 2008.

[25] L. Y. Lee, Y. S. Lin, & C. P. Chu. *Enhancement of personal concept map constructing for effective assessment*. In Teaching, Assessment and Learning for Engineering (TALE), 2012 IEEE International Conference on (pp. W1A-1). IEEE. 2012. DOI: 10.1109/tale.2012.6360395.

[26] M. de la Villa, F. Aparicio, M. J. Maña, & M. de Buenaga. *A learning support tool with clinical cases based on concept maps and medical entity recognition*. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (pp. 61-70). ACM, 2012. DOI: 10.1145/2166966.2166978

[27] M. Al-Sarem, M. Bellafkih, & M. Ramdeni. *An approach for mining concepts' relationships based on historical assessment records*. Procedia Engineering, 15, 3245-3249, 2011. DOI: 10.1016/j.proeng.2011.08.609.

[28] M. Elhoseiny, & A. Elgammal. English2mindmap: *An automated system for mindmap generation from english text.* In Multimedia (ISM), 2012 IEEE International Symposium on (pp. 326-331). IEEE. 2012. DOI: 10.1109/ism.2012.103.

[29] M. Kantardzic, H. Hamdan, & B. Djulbegovic. *Artificial Neural Networks (ANN) Approach in Diagnostics of Polycythemia Vera*. International Journal of Computers and their Applications, 8, 74-79. 2001.

[30] N. S. Chen, C. W. Wei, & H. J. Chen. *Mining e-Learning domain concept map from academic articles*. Computers & Education, 50(3), 1009-1021. 2008. DOI: 10.1016/j.compedu.2006.10.001.

[31] N. S. Chen, P. Kinshuk, C. W. Wei, & H. J. Chen. *Mining e-learning domain concept map from academic articles*. In Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06) (pp. 694-698). IEEE, 2006. DOI: 10.1109/icalt.2006.1652537

[32] N. Yi, & H. Li. *A practical approach for automatically constructing concept map in E-learning environments.* In Progress in Informatics and Computing (PIC), 2014 International Conference on (pp. 582-586). IEEE. 2014. DOI: 10.1109/pic.2014.6972401

[33] P. Pirnay-Dummer, & D. Ifenthaler. *Reading guided by automated graphical representations: How model-based text visualizations facilitate learning in reading comprehension tasks*. Instructional Science, 901-919. 2011.

DOI: 10.1007/s11251-010-9153-2

[34] R. Richardson, & E. A. Fox . *Using concept maps in digital libraries as a cross-language resource discovery tool.* In Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries (pp. 256-257). ACM, 2005. DOI: 10.1145/1065385.1065443.

[35] R. Y. Lau, D. Song, Y. Li, T. C. Cheung, & J. X. Hao. *Toward a fuzzy domain ontology extraction method for adaptive e-learning*. IEEE Transactions on Knowledge and Data Engineering, 21(6), 800-813. 2009. DOI: 10.1109/tkde.2008.137

[36] R. Mitkov. *Anaphora resolution: the state of the art.* School of Languages and European Studies, University of Wolverhampton. 1999.

[37] R. Feldman, & J. Sanger . *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press. 2007. DOI: 10.1017/cbo9780511546914

[38] R. Tavares. "*Aprendizagem significativa, codificação dual e objetos de aprendizagem*." Revista Brasileira de Informática na Educação 18.2. 2010. DOI: 10.5753/rbie.2010.18.02.04

[39] S. S. Tseng, P. C. Sue, J. M. Su, J. F. Weng, & W. N. Tsai. *A new approach for constructing the concept map*. Computers & Education, 49(3), 691-707. 2007. DOI: 10.1016/j.compedu.2005.11.020

[40] S. M. Chen, & P. J. Sue. *Constructing concept maps for adaptive learning systems based on data mining techniques*. Expert Systems with Applications, 40(7), 2746-2755, 2013. DOI: 10.1016/j.eswa.2012.11.018

[41] S. Lee, Y. Park, & W. C Yoon. *Burst analysis for automatic concept map creation with a single document*. Expert Systems With Applications, 42(22), 8817-8829, 2015.

[42] S. M. Bai, & S. M. Chen. *Automatically constructing concept maps based on fuzzy rules for adapting learning systems*. Expert systems with Applications, 35(1), 41-49. 2008. DOI: 10.1016/j.eswa.2007.06.013

[43] S. Wang, & L. Liu. *Prerequisite Concept Maps Extraction for Automatic Assessment*. In Proceedings of the 25th International Conference Companion on World Wide Web (pp. 519-521). International World Wide Web Conferences Steering Committee, 2016. DOI: 10.1145/2872518.2890463

[44] S. M. Bai and S. M. Bai. *A new method for automatically constructing concept maps based on data mining techniques*. In: 7th International Conference on Machine Learning and Cybernetics, ICMLC. p. 3078-3083. 2008. DOI: 10.1109/icmlc.2008.4620937

[45] T. B. S. Gava, C. S. de Menezes, and D. Cury. "*Aplicações de mapas conceituais na educação como ferramenta metacognitiva*." III International Conference on Engineering and Computer Education-ICECE 2003. 2003.

[46] T. Strzalkowski. Natural language information retrieval (Vol. 7). Springer Science & Business Media. 1999. DOI: 10.1007/978-94-017-2388-6

[47] I. Vekiri. What is the value of graphical displays in learning?. Educational Psychology Review, v. 14, n. 3, p. 261-312, 2002.

[48] Y.F. Le Coadic. *A ciência da informação*. Briquet de lemos Livros, 1996.

[49] W. M. Wang, C. F. Cheung, W. B. Lee, & S. K. Kwok. *Mining knowledge from natural language texts using fuzzy associated concept mapping*. Information Processing & Management, 44(5), 1707-1719. 2008. DOI: 10.1016/j.ipm.2008.05.002

[50] G. Li, B. C. Ooi, J. Feng, J. Wang, & L. Zhou. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 903-914). ACM. 2008.

[51] E. K. Jacob. Classification and categorization: a difference that makes a difference. Library trends, 52(3), 515. 2004.

[52] T. T. Gruber et al. A translation approach to portable ontology specifications. Knowledge acquisition, v. 5, n. 2, p. 199-220, 1993. DOI: 10.1006/knac.1993.1008