PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTERY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY OF MOHAMED BOUDIAF - M'SILA

FACULTY: Mathematics and Computer Science
DEPARTMENT: Computer Science
N°:……

DOMAIN: Mathematics and Computer Science
BRANCH: Computer Science
OPTION: Information and Communication Technology

A Dissertation in Fulfillment
For the Requirements of the Degree of Master

By: Hamza Benkhelil

SUBJECT

Automatic generation of concept map from text

Defended publicly on: ../06/2017

Board of Examiners:

……	University of M'sila	Chairman
Mr. Said Hamani	University of M'sila	Supervisor
……..	University of M'sila	Examiner

Academic Year: 2016/2017

Table of Contents

List of Figures

List of Tables

INTRODUCTION

Concept maps (CMs) are graphical tools for organizing and representing knowledge [1]. They include concepts, usually represented by circles or boxes, connected by directed edges to form relationships, which are represented by annotated lines between the concepts, Figure 1.1 illustrates an example of a concept map that describes 'what are birds ?' [2]. A concept can be a noun or a short noun phrase, and the relationship label is usually presented as a verb or verb phrase. According to the example in Figure 1.1, 'birds', 'feathers', and 'eggs' represent concepts and connecting words such as 'lay', 'have' represent relationship labels. The concept-relation-concept triple forms a proposition, which represents a meaningful statement to interpret, in the example concept map, '[birds]-[lay]-[eggs]' and '[feathers]-[help to]-[fly]' triples form propositions.

Another characteristic of concept maps is that the concepts are represented in a hierarchical fashion with the most inclusive, most general concepts at the top of the map and the more specific, less general concepts arranged hierarchically below [1]. Another important characteristic of concept maps is the inclusion of cross-links. These are relationships or links between concepts in different segments or domains of the concept map. Cross-links help us see how a concept in one domain of knowledge represented on the map is related to a concept in another domain shown on the map, facilitating creative thinking [3]. For instance, 'high metabolism provides energy' represents a cross-link that creates interconnections between the concepts 'rapid digestive systems' and 'food'.

Many educators and researchers have exploited CMs in a number of ways, including evaluation or assessment tools [4, 5], cooperative meaningful learning tools in education [6, 7], and advance organizer and visualization tools [8]. Also, CMs have been used as a means for communicating information for organizing ideas and promoting problem solving strategies [9, 10]. A CM can be used as an intermediate step in ontology learning for a particular domain to support the acquisition of domain knowledge [11].

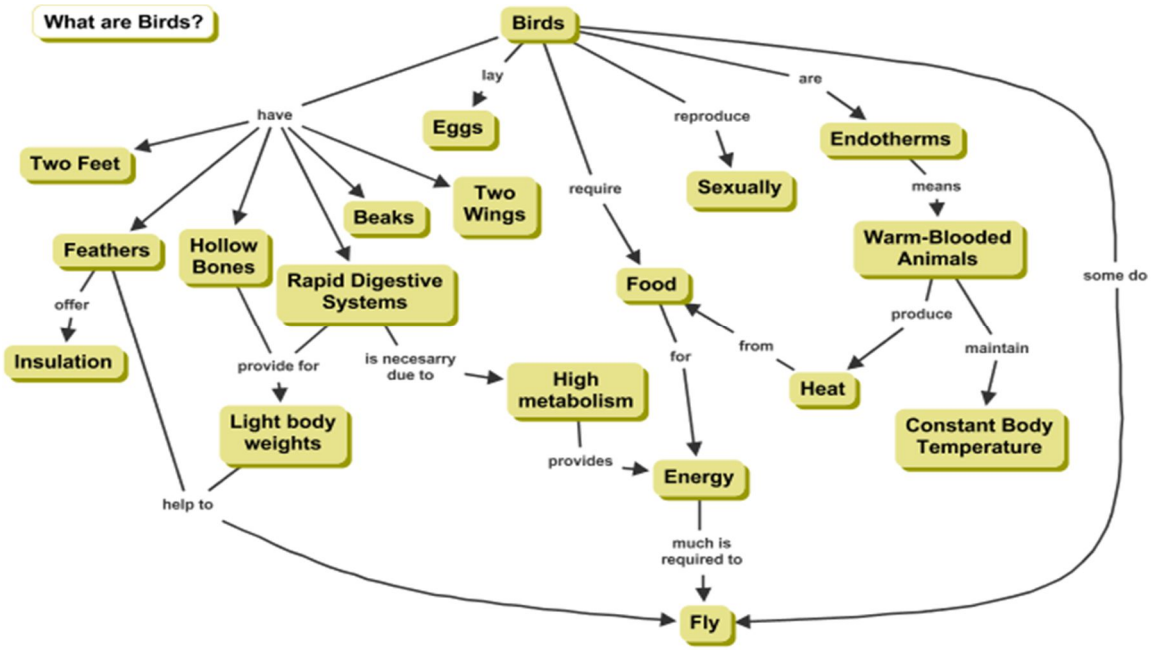

Figure 1.1 An example concept map [2].

The construction of concept maps is typically done either manually or automatically. When the concept maps are created manually, the maps are most often used to represent how a user or users understand a topic or a domain of relation concepts. The construction of such concept maps is a process of learning and discovery where users add more concepts and links to the maps as they learn and understand more about their studying topics or domains. When the maps are constructed automatically, the maps are generated from a body of documents related to the topics or domains that the concept maps represent. The construction process is focused on computational learning, i.e., using computer's learning algorithms to extract and learn about concepts and their relationships from the underlying documents and represent them in a graphical form. While there have been significant numbers of research on using manually constructed concept maps for teaching and learning [12, 13], there is little research on applying automatically constructed concept maps for concept learning and exploration. It is important to recognize differences between the manually and automatically constructed concept maps and develop different applications for them.

Therefore, the objective of this thesis is to investigate computational approaches applied in construction of concept maps from text and develop new method.

.

1 Motivation

For educational purposes, concept maps are used as a learning tool for the students. An effective concept map can be considered as a map that is easily understood by a second party. Generating an effective concept maps is sometimes considered as a complex task as users may find it difficult to remember some concepts of a certain topic hence the need for automatic generation of concept map.

The motivation for concept map generation from text arises from three key supporting reasons;

1 Concept maps are a highly effective educational tool, grounded in important learning theories [1, 3].

2 The need to reduce the issues associated with manual construction of concept maps by learners or construction of expert maps[1] [2].

3 Analysis and understanding of text research [14].

2 Organization of the thesis

The organization of this thesis is as follows: we begin in Chapter 1 by presenting the notion of concept maps, their uses and how they are constructed. Chapter 2 provides a detailed review related to concept map mining from various text sources. Chapter 3 we introduce our method of generating concept maps from text.

[1] Expert maps are identified as concept maps which are usually constructed by human experts within a particular domain or topic.

CHAPTER 1

CONCEPT MAPPING: A REVIEW

This chapter includes an overview of concept mapping. Concept map uses and applications are also briefly reviewed.

1 Overview of Concept Mapping

In 1972, Joseph Novak and his team of researchers at Cornell University were studying children's emerging understanding of science concepts. They created 28 science lessons, and attempted to understand how children developed knowledge of the concepts presented to them. While analyzing the many interviews they had held with the participants, they found it particularly difficult to determine if the children had acquired new understanding of concepts, and whether this understanding was integrated into their existing knowledge framework [15].

The integration of new concepts into a learners existing knowledge framework was based on psychologist David Ausubel's assimilation theory. Ausubel makes the distinction between learning meaningfully and by repetition, or rote. In learning meaningfully, learners assimilate new concepts into their existing concept and proposition framework, which Ausubel refers to as the learner's cognitive structure. He describes the cognitive structure as a hierarchical structure, with general concepts at the top, and more specific concepts underneath. He describes two processes: the process of subsumption, in which learners subsume new knowledge in existing concepts and propositions, and superordinate learning, in which prior knowledge is subsumed into new more general or abstract concepts [16].

Because of the trouble Novak and his colleagues had determining if the children had learned the meaning of the taught science concepts; they searched for a tool that would facilitate explicitly displaying the children's cognitive structure. As such, they developed the Conceptual Map. Since then, the concept map has proven itself useful in many different applications.

The concept map, as formalised by Novak and his team, is a structured diagram containing concepts connected by linking phrases. It's a hierarchical tree-like structure, which is often concentrated around a focus question. The focus question creates the context and helps

determine the scope of the knowledge that is to be represented. At the top of the concept map hierarchy the superordinate concept is displayed. Concepts can be abstract terms referring to an object or event, with its meaning in part tied to the concepts directly related to it. Novak defines concepts as: "a perceived regularity in events or objects, or records of events or objects, designated by a label" [15]. Concepts are indicated by a label inside a box, often consisting of a noun phrase. A different type of abstraction is a construct. Constructs are also concepts; however these are more often intangible or inferential. As an example, motivation is a construct that is not directly observable by itself, it is inferred by other concepts related to it. Although constructs are not displayed differently form concepts, they are often higher up in the concept maps hierarchy.

Linking phrases are the arcs connection associated concepts, in most cases consisting of a verb phrase. A linking phrase is inherently bi-directional: a concept linking from concept A to concept B with the linking phrase has-child, also links concept B to concept A with the linking phrase has-parent. Concepts connected by a linking phrase are referred to by Novak and Canas [1] as propositions: "Propositions are statements about some object or event in the universe, either naturally occurring or constructed. Propositions contain two or more concepts connected using linking words or phrases to form a meaningful statement".

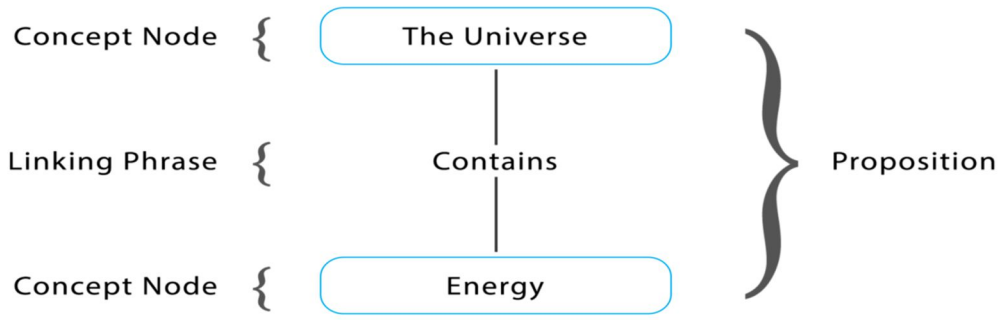

Figure 2.1 Example of a proposition [2].

Another feature of concept maps are cross-links. Cross-links are usually realized after the construction of the initial concept map. Cross links allow two concepts in different parts of the domain to be linked together and illustrate how these may be related to each other. Research by Novak and Gowin indicates that student's identifying cross-links have reached a high level of understanding of the domain [17]. Figure 2.2 below displays a conceptual map as defined by Novak and canas [15].

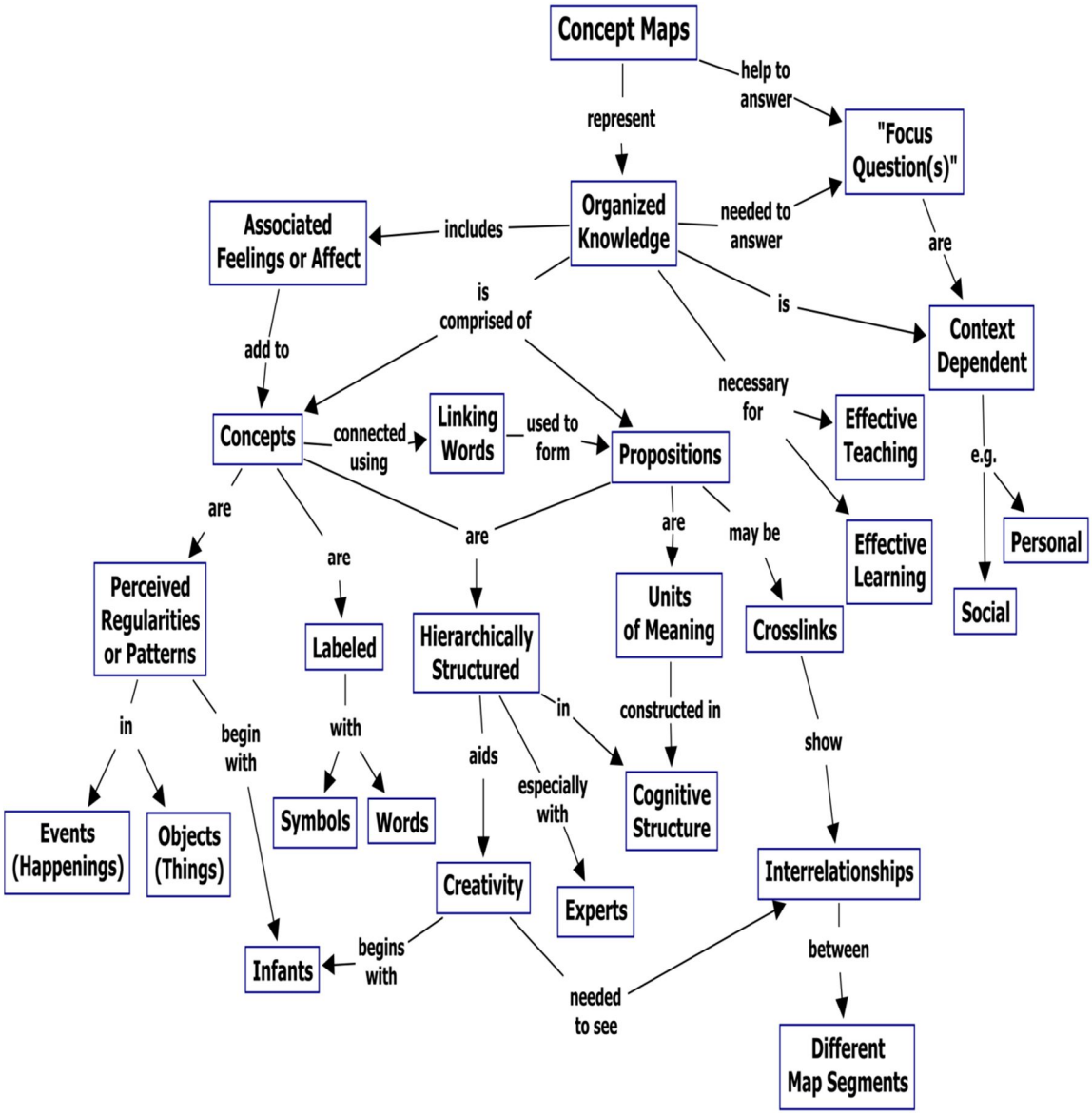

Figure 2.2 A conceptual map as defined by Novak and canas [1].

Although the concept map has proven to be effective tools for educators, it is not without its drawbacks. Several restrictions to the application of concept mapping apply. Firstly, it is not a simple and quick visualization tool because of the formal rules that have to be abided by. It can be difficult to properly identify the concepts and the relationship between them. Secondly, domains with particularly large interconnectedness between concepts may become very difficult to model. Visually complex concept maps may feel chaotic and its interpreters could become overwhelmed attempting to understand it fully [18].

2 Constructing a Concept Map

According to Novak and Canas [1], concept maps often target a focus question. The focus question serves as a starting point. Then, a "parking lot" of concepts related to the focus question is made. This serves as a preprocessing step to get a clear idea of the domain. The concepts are then put in hierarchical order of importance. Afterwards, the concepts are placed on the map one by one, while simultaneously identifying the links between them. After the links have been drawn, they are given names as to represent the relation between them. Finally, when the map is considered to be sufficiently covered by concepts, cross-links can be added to link different areas of the domain.

3 Applications and uses of Concept Maps

Concept maps have been widely used in education. They have been demonstrated to be a successful instructional tool to help learners in their understanding process. Concept maps are popular as they aid in creative thinking, knowledge extraction, planning, note taking, summarization [19], idea generation, and knowledge creation [20] and as assessment [21] and evaluation tools [22]. Concept maps can also be used to summarize papers. According to Richardson and Fox [23], a concept map can be as good a summary as an abstract, and are easier to automatically prepare and translate than a written abstract.

Darmofal [24] has used concept maps and concept questions for engineering university level to help in their conceptual understanding of the discipline and stimulate thinking. In [25] concept maps have been used in searching through historical archives. These maps provide a representation of the important retrieved entities, which might be used in later searches. Maria [26] demonstrated the application of concept maps in conjunction with practical and cognitive apprenticeships to teach and improve programming skills in holistic learners. The use of concept maps proved to stimulate meaningful learning in undergraduate medical students taking a PBL (problem-based learning) [27]. McClure [28] researched on the use of concept maps to assess learner's knowledge on certain concepts.

The use of concept maps is not restricted to education, but they are used in business planning. Public administration and health sector, among others. Concept maps have been employed in community mental health [29] for program planning and evaluation purposes.

Compared to other knowledge elicitation tools, concept mapping is considered as an efficient method for generation models of domain knowledge [30]. When integrated with other systems, concept maps have been used as interfaces for intelligent software (i.e. knowledge based systems and tutoring systems) in various domains [31].

From an educational instructor's point of view, concept maps can be used to reveal a learners' understanding or misconception [32] of a certain knowledge domain. There are no "correct" concept maps but often the teacher's concept map is used as a reference map [33]. However, a teacher's map reflects the teacher's way of thinking. For a more objective map, a different approach used to construct the concept map is applied. . Automatically generated concept maps are less biased, easy to generate and can be used as reference maps. Hideo [34] developed a concept mapping software that "supports the externalization of ideas, reflection on thinking processes and dialogues" by allowing collaborative learning by permitting several users to construct one concept map. There have been several tools like CmapTools [35]. Clouds [36], Leximancer [37] and GNOSIS [38], which attempt to construct concept maps, in interaction with the users to generate concept maps automatically.

4 Summary

In summary, a concept map is a type of knowledge representation to develop mental schemas or mind maps that act as a reference for future actions and thinking [39]. Concept maps can be applied in different areas and not limited to the education field. A common issue arising is the difficulty in evaluating different concept maps. Not even a human expert can say for certain what a correct concept map should look like. Therefore, it can be hypothesized that an automatically generated concept map has a reduced degree of bias compared to a manually generated concept map.

CHAPTER 2

CONCEPT MAP MINING: A REVIEW

In this chapter, we present an overview of concept map mining (CMM) process and discussing some of the existing approaches related to CMM. We will then make a comparison of the different approaches discussed.

1 Overview of Concept Map Mining

According to the definition by Villalon [40], the concept map mining (CMM) process can be expressed as the proper extraction of a concept from a document (D). This process has three steps:

1. Concept extraction (CE) and identifying the set of concepts (C)
2. Relation extraction (RE) and identifying the set of relations (R)
3. Topology extraction (TE) and identifying a generalization (G) of the set of concept

CE must be the first step of the process, because C defines R and G. Every term used to describe a concept or a relationship must appear in the document, therefore D itself defines all potential words (or phrases). This idea can be formalized by defining a document as a triplet $D = C_d, R_d, G_d$ where C_d corresponds to all the concepts, R_d corresponds to all the propositions, and G_d corresponds to the levels of generalization expressed in the document.

The next stage is to identify the subsets *C, R* and *G* that are a good summary of D. Figure 3.1 illustrates the concept map mining process.

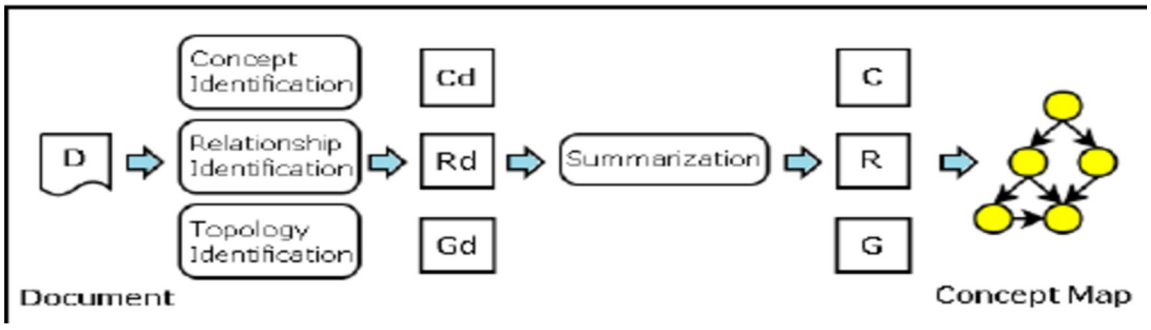

Figure 3.1 Concept map mining process [40].

The goal of the CMM process is to produce CM that is an accurate visual abstract of a source text. Created map is intended for human analysis, and it should not contain too many concepts, preferably 15-25 [1]. In educational context, the terminology used in a document is important for users, so the CM should be represented using terms that the author used in the original text.

The source of the CMM technique can be traced back to the early work of Trochim, who proposed a concept mapping process that combines a group activity with statistical analyses [42]. The group of participants during brainstorming session creates a set of statements relevant to the domain of interest. Each participant sorts and rates every statement, creating individual similarity matrix. All personal matrices are summed together into a group proximity array. The most important statements are chosen using a multidimensional scaling (MDS) and hierarchical cluster analysis. This approach, based on weight calculation, statistical and data mining techniques is still commonly used in many contemporary CMM methods [41].

2 Concept Map Mining Methods and Approaches

A number of studies have focused on the automatic generating of concept maps, or similar representations, for various applications. Some researchers follow the strict definition of a hierarchical CM, while others use knowledge representations that are somewhat more variable. In the beginning of this section, we describe the different methods used in the natural language processing (NLP) field for CMM. In following sub-suctions, a short overview of main approaches used in current CMM studies is given.

2.1 Methods used Concept Map Mining[2]

A CMM process can be carried out by NLP methods used in tasks such as information extraction (IE), information retrieval (IR) and automatic summarization. IE is the process of automatic extraction of structured information, such as entities and relations, from unstructured textual sources. IR is area concerned with searching for information in documents and metadata about documents, while the goal of automatic summarization is to distil content from a source, and present the most important content to the user in a condensed form [75].

[2] This section id based on the review of (Zubrinic et al.)[41]

Methods traditionally used in these areas are rule-based statistical and machine learning methods. More recently, there has been interest in combining finite-state machines with conditional-probability models, like maximum entropy Markov models and conditional random fields [43]. Most of classical summarization methods are likewise numerical, and based on a weighting model where system weights text elements according to simple word or sentence features, or statistical significance metrics like term frequency-inverse document frequency (TF-IDF). Machine learning methods often provide accurate extraction based on classification, using binary or fuzzy logic. Such method can be used as a main method, or in hybrid systems to supply resources to other processes. Contemporary approaches include hybrid techniques with usage of algorithms in combination with third-party datasets [44], [45], summarization based on fuzzy logic and swam intelligence [46].

In NLP field, numerical methods can be enriched with dictionaries of terms or linguistic tools and techniques [43]. Problem with dictionaries arises from the fact that dictionary is usually an external resource. It has to be previously created for specific domain and it requires further operations for handling new content. A limiting factor for use of linguistic techniques is that appropriate tools and methods are not available for many languages [41].

2.2 Concept Map Mining Approaches

The data sources used for the concept map construction in the approaches is in most cases dependent on the purpose of the research. Broadly, we can identify structured and unstructured data sources.

2.2.1 Mining from unstructured textual data sources

Unstructured text is considered to be regular text that is not pre-annotated either computationally or manually. Considering the number of documents used for one CM creation, there are two groups of techniques. The first group contains techniques that create one CM from a single document [47], [48]. Multiple documents are used as a source in the second group of techniques [49], [50].

The goal of most studies in this area is to produce a starting CM model, which can speed up the process of CM creation for later refinement by a person, or by another automatic process. Some of created maps are fully completed and contain concepts connected with labelled

relationships [51], [53]-[48], [54]-[55]. Other researches created CMs with connected concepts, but without labelled relationships [47], [56], [49], [57], [58]-[59], or extracted only concepts [60].

The statistical approaches either make use of a quantitative data source or quantify the contents of the data source. In the quantitative data source, a relation is established between statistics and concepts, and further analyzed to extract regularities. The quantification of the data source involves counting the appearance of concepts in text [18]. These methods are commonly used in combination with methods such as machine learning or linguistic. Although the results of the approach do not represent a conceptual map as formalized by Novak and Canas [1], one can consider the results to be as scaffolds for exploration of the source document content. The most used statistical methods are analysis of co-occurrences between terms and different term frequency analyzing techniques such as term frequency, inverse document frequency (TF-IDF) and Latent Semantic Analysis (LSA) [41].

Research by Clariana and Koul [47] defined a list of important terms in the Biological domain. The co-occurrence of this pre-defined list was compared with the 'terms' in students written summaries with less than 30 sentences and constructed an 'aggregated proximity array' by assigning '1' when co-occurred and '0' otherwise.

Concept map mined from Chinese 'news articles' reflected the scientific knowledge contained in daily news stories [57]. This approach utilized a rule-based algorithm to extract key terms. Their algorithm was applied to a collection of 100 Taiwan news articles and obtained 10% improvement in the 'accuracy' compared to the baseline approach. The first k terms (k =30) were ranked based on the TF-IDF measure and the association between the top k terms were calculated using co-occurrence analysis of nearly 25,000 Chinese news articles. Five assessors were recruited to judge the relevancy between selected 30 term pairs. Based on their judgement, 69% of the associated terms were relevant within the selected set of terms. This approach also lacked 'relation labels' and hierarchy of concept maps.

In general, statistical methods such as TF-IDF and co-occurrence analysis do not rely on domain knowledge or external resources. Therefore, statistical methods can be applied to any domain for knowledge extraction. However, statistical methods commonly suffer from probable semantic loss [41]. These methods fail to identify and overcome semantic ambiguities like pronouns and determiners which are used as a substitute for a noun or noun phrase. Additionally, statistical methods are not able to detect synonyms (i.e. word with the similar meaning of another

word) and homonyms (i.e. words that share the same spelling and pronunciation but have different meanings). Instead, statistical methods focus on the distribution of information within the corpus [61].

Alternatively, the linguistic analysis approach deals with common language processing techniques, such as sentence boundary detection, tokenization, part of speech tagging and chunking. In some cases, external assets such as WordNet [72] or NomBank are used to compliment the performance of the system by finding morphological variations of words [62], [63]. The problem with linguistic techniques is that they are limited to a specific language, and linguistic resources must exist for used language. The majority of linguistic tools and methods are based on the English language, and researchers mostly use them in CMM [14], [64], [65], [66], [54]-[55], [69].

In order to avoid problems of language techniques, in his PhD thesis Richardson [64] described a method for automatic translation of a CM from one language to another. As practical example, he created maps from theses and dissertations in computing in the English language and translated them to Spanish. During a CMM process, syntactic dependencies between noun phrases were recognized using WordNet's morphological functions, and translated into CM propositions. Smaller CMs were summarized using simple heuristics, taking top 20 concepts. For summarization of huge CMs with more than 100 concepts, the value of TF-IDF indicator was used. Created CMs were translated into Spanish using an algorithm for translation of individual words and shorter phrases based on bilingual dictionary.

Willis and Miertschin [48] used centering resonance analysis for creation of CMs used in the process of students' assessment. Using linguistics theory, that method creates word network of nouns and noun phrases in order to represent main concepts, their influence, and their interrelationships.

In general, linguistic methods extract nouns (or compound nouns) as concepts and verbs as relations. However, there can be nouns existing which are not concepts in that particular domain. Additionally, there can be verbs which act as nouns in some domains (e.g. gerund verbs).

Machine learning methods can be of several forms such as supervised learning and unsupervised learning. Classification systems such as a naïve Bayes classifier [11], association rules such as fuzzy rules [67], [68] or clustering techniques to extract concept maps from text

sources are the techniques most commonly used in this process. The majority of relation extraction approaches utilize machine learning methods.

The classification technique is used for the extraction of key terms in the TEXCOMON [11] software tool. This application uses a simple Kea algorithm based on a naïve Bayes classifier. Lee [70] uses an a priori algorithm and association rules for the automatic construction of a CM from learners' wrong answers to exam questions. As the method creates only association rules based on questions that incorrectly answered, it can miss information associated with correctly answered questions. An improved method Chen and Bai [67] creates association rules for all questions and achieves results that are more appropriate for use in adaptive learning systems.

A number of methods implement fuzzy reasoning and fuzzy techniques. A hybrid algorithm called “concept frame graph” Rajaraman and Tan [55] uses grammar analysis and fuzzy clustering techniques for the creation of a CM knowledge base that is represented using concept frames. Each concept frame consists of information about a concept: its name, context, set of synonyms and its relationship to other concepts.

CMM methods that use fuzzy association rules to extract predicates from a learner' historical testing records Bai & Chen [58], Sue [59] are used in adaptive learning environments. In particular, they use a fuzzy association method to search for non-explicit links that exist among concepts. This method employs concept weights combined with fuzzy heuristic. Similar approaches create CMs from newspaper articles [66], and messages posted to online discussion forums [69].

Generally, in the most case the machine learning techniques are not used in the CMM alone, but used to complement performance of the linguistic processes (known as ‘hybrid’ methods). Linguistic or machine learning methods for knowledge extraction and statistical methods for ranking the extracted knowledge [68], [63], [62], [65]-[71]. In addition, some other works utilized linguistics methods for knowledge extraction while machine learning methods to build the concept maps [14]. As a result, the output of the two approaches combined area of significantly better quality than the other approaches [18].

TextStorm extracted relations between concepts as binary predicates (e.g. John eats meat => eat(John, meat)) using Natural Language Processing techniques and the Cloud system interactively completed the missing knowledge in concept maps by asking primitive questions

from users (e.g. Question: Define lions with the predicate 'is-a' => Answer: is-a (lion, animal)) using machine learning methods [14]. With a sample of 21 small text files of articles, manuals, educative texts, TextStorm achieved a correctness mean of 52% in extracting binary predicates. Although the approach operated independently from domain knowledge, TextStorm limits its usage since it depended on WordNet [72] for lexical verifications of the words. Further, this approach is limited to affirmative and declarative sentences.

The Fuzzy Association Concept Mapping (FACM) technique has been proposed to automatically extract concept maps from abstracted short texts [66]. This technique utilised linguistic methods to extract propositions (in the form of 'concept-relation-concept' triples) from text and interactively refined the extracted proposition with the use of human recommendations using fuzzy set theory. This framework was evaluated with the Science Citation Index (SCI) abstract database and CNET news. The proposed FACM approach was compared with a baseline algorithm called ANNIE (A Nearly-New Information Retrieval System) based on the GATE platform [73] and obtained higher precision (87% for 'abstracts' and 83% for 'news articles') and higher recall (78% for 'abstracts' and 74% for 'news articles').

Valerio and Leake [62] studied the effect of auto-generated concept maps for problem solving. Their study measured the reading comprehension skills in terms of both 'speed' (i.e. time taken to answer question) and 'accuracy' in answering questions. A group of 16 undergraduates and graduates were provided with 60 questions to answer using a text document, a concept map constructed manually or auto-generated maps. Results showed that providing auto-generated concept maps improved user speed in answering questions, for well-written documents whose size enabled generating a single concept map with a limit of 30 most important concepts. However, there was no significant difference between each resource on 'accuracy'.

Concept map mining from a Biology text book [63] utilised the SemNet formulations [74] to generate concept maps. In this approach, key terms were used as 'start nodes' of a triple, where the 'end node' could be either key terms or a complete proposition. The concept maps generated, which was consistent with SemNet, allowed comparison with thousands of Biological triples available online. They reused key terms existing in the 'glossary' and 'index' of text books and the test-prep study guides. After applying triple extraction algorithms to extract nearly 4400 triples, a manual categorisation was carried out to cluster them according to relationship types. Finally, statistical-based filters were applied to discard triples that were not suitable for concept

map exercise generation. With the use of two experts having background in Biology and pedagogy, a Wilcoxon signed ranked test, pairing human and computer generated maps found that computer generated maps were more 'accurate' to be utilised as expert maps ($Z = 2.13$, $p < .03$). One of the benefits of having a 'text book' as the mining source is that text books contain grammatically complete sentences with minimal ambiguities (e.g. pronouns). This approach had limitations in that the auto-generated concept maps contained fewer links (approximately 3.5 times) than expert maps. Therefore, it was difficult to reuse them to compare with student constructed or modified maps. Additionally, this system failed to extract every triple from every sentence which resulted in a low recall.

Concept map mining from students' written essays on the topic of 'English as a global language' helped visualise the concepts and relations included in essay form [65]. This approach utilised grammar trees to identify concepts and Latent Semantic Analysis (LSA) for ranking. The system suggested a list of concepts from student essays (approximately 12 concepts per essay) and human annotators built concept maps from them, enabling the visualisation of the essay as a concept map. Even though the human-machine agreement (i.e. accuracy) was not greater than inter-rater agreement, the system reported promising results when compared to the related works in the literature. Inter-rater agreement is the agreement between human evaluators when more than one evaluator is involved in an evaluation task [75].

The hybrid method is intended to overcome the specific issues discussed under statistical and linguistic methods. Statistical methods suffer from probable semantic loss while linguistic methods are suited for well-written natural language text [40]. Additionally, linguistic methods extensively rely on external databases such as WordNet [72].

2.2.2 Mining from structured textual data sources

Data sources in the structured category can be defined as data that was in some form annotated [75]. Ontologies are used as structured textual data sources in CMM. Their usage in computing has increased during last decade, especially after introduction of the semantic Web. In the context of knowledge sharing, an ontology is a description of objects and relationships between them [76].

CMs and ontologies are quite similar an ontology can be formalized as a triple subject-predicate-object, and a CM as concept-relation-concept. Both of them consist of classes (or

concepts) and relationships among them. Unlike CMs, ontologies are more formal and more expressive because of their attributes, values and restrictions. The basic approach used for translating an ontology to a CM is through a direct mapping of ontology classes and associations into concepts and relationships [41].

Kim et al. [52] proposed a way for translation of ontology of English vocabulary into a customized CM. The translation is processed by the software agent that directly maps ontology classes and properties to CM propositions. The algorithm described in [77] follows the same approach. The first part of that algorithm searches for ontology class hierarchy and discovers instances of classes. Synonyms, intersections and unions among classes are translated into links between concepts. Equally, all properties, data types and values became new concepts. At the end, the algorithm checks and corrects symmetric and transitive links.

3 Comparisons

Table 3.1 compares the CMM systems discussed above based on underlying data source, methods and the features in the form of relation and hierarchy extraction. Each of the CMM systems was capable of extracting concepts.

Systems	Data source	Method	Relation	Hierarchy
TextStorm [14]	Text file	Hybrid (linguistic and machine learning)	Yes	No
Chen et al., [49]	Academic articles	Statistical (term frequency and relationship strength)	Yes	No
Chen S.-M. and Bai, [67]	Answers to questions (both correct and incorrect)	Machine learning (association rules)	Yes	No
ALA-Reader [47]	Written summaries of students	Statistical (co-occurrence)	Yes	No
Lau et al., [68]	Discussion forums	Hybrid approach (lexico-syntactic patterns and statistical method)	Yes	Yes (fuzzy taxonomy of relations)

Lau et al., [70]	Incorrect answers to exam questions	Machine learning (association rules)	Yes	No
Olney [63]	Text book	Hybrid (natural language processing and statistical methods)	Yes	No
TP-CMC [77]	Learner's historical testing records	Machine learning (Fuzzy association rules)	Yes	No
Tseng Y.-H. [57]	Chinese news articles	Statistical method (rule-based extraction algorithm & co-occurrence analysis)	Yes	No
Valerio and Leake [62]	Well-written text document	Hybrid method (linguistic and text mining)	Yes	Yes
Villalon and Calvo [65], [71]	Student essay	Hybrid method (linguistic and latent semantic analysis)	Yes, in the later work [74]	Creates manually
TEXTCOMON [11]	Text documents	Hybrid (machine learning and linguistic method)	Yes	No
Wang et al. [66]	Abstracted short text	Hybrid (linguistic and fuzzy set theory)	Yes (interactively built with human users)	No

Table 3.1 Comparison of concept map mining systems [61].

4 Summary

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

A frequent problem that occurs in the recognition of relationships, if finding to which noun phrase a pronoun phrase refers. Problem is known as anaphora resolution. Without proper

anaphora resolution, much of semantic information from a text could be lost. Determining a link label and direction is an additional difficulty [41].

In the next chapter, our CMM procedure for automatic generation of concept maps from text is proposed and described in detail.

CHAPTER 3

CONCEPT MAP MINING FROM TEXT

In this chapter, we describe our method for generating a concept map from text.

The Concept Map Mining from text was developed in JAVA and adopted Stanford Core NLP tools for the task of text preprocessing and triple extraction from sentences. Stanford Core NLP library is commonly used as research tool. The Stanford part-of-speech tagger is accurate up to 97% to identify part-of-speech tags of natural English texts [78].

We used a XML-based file format called Concept Map Extensible Language (CXL) to store the propositions and visualized the concept map using IHMC CmapTools [79].

1 Description of Our Method

Our method combines unsupervised algorithm and NLP tools. The process for creating concept map consists of five stages as shown in figure 4.1.

- **Preprocessing:** this stage is responsible for providing all the resources needed for performing the extraction process. Most of these steps are implemented using Stanford CoreNLP.
- **Concepts extraction:** this stage involves the extraction of potential candidates for concepts in a CM.
- **Creation of proposition:** in this stage, we search for relationships between concepts.
- **Ranking:** in this stage, the propositions are ranking based on their concepts importance.
- **Visualisation:** in this stage, the propositions are translated into the CXL format and visualised using IHMC CmapTools [79].

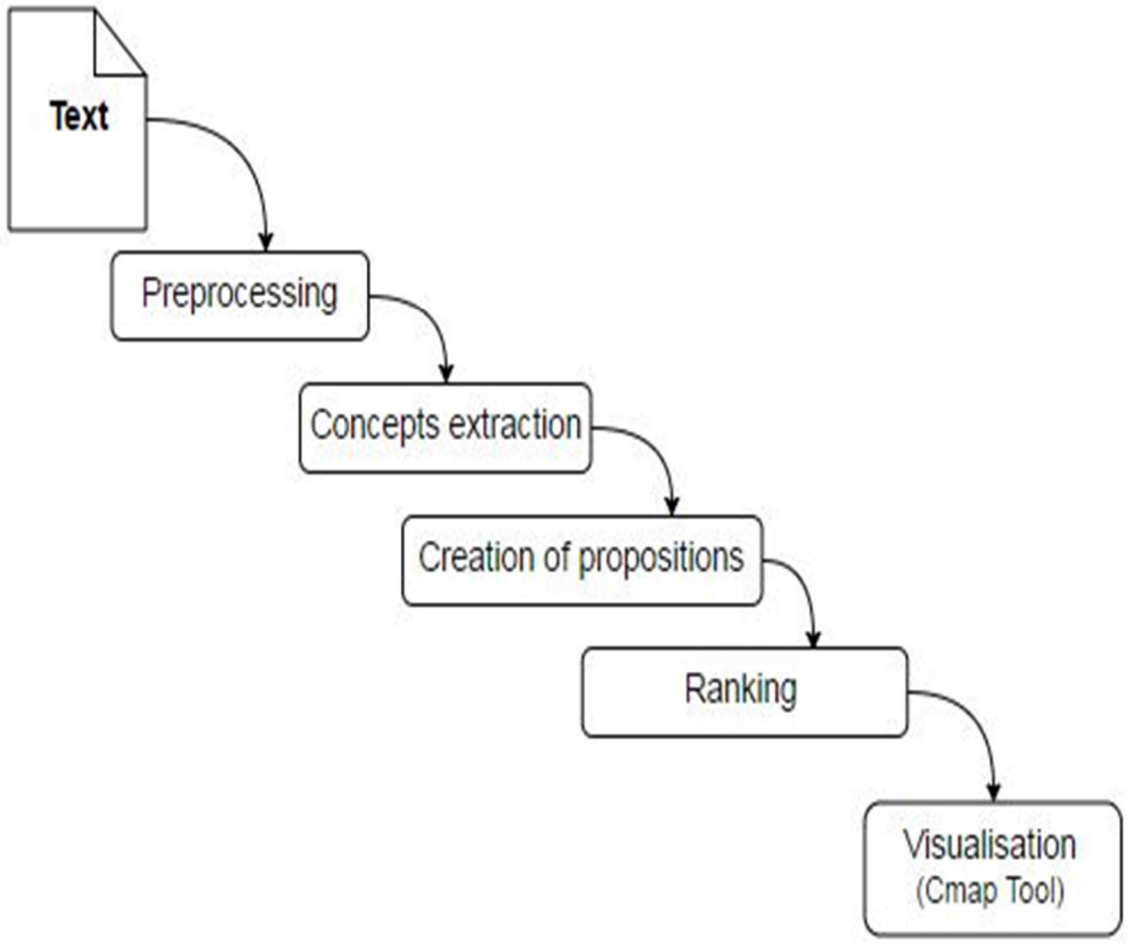

Figure 4.1 The process of generating concept map from text.

1.1 Preprocessing

This stage includes normalization, tokenization and morphological analysis, text segmentation and syntactic analysis.

The normalisation step modifies the data source to ease further processing. This comprises removing special characters, for instance ($, #, &) and lemmatisation.

The steps for tokenization and morphological analysis are focusing on individual terms. The first step divides the text into tokens; the second identifies the part of speech of each identified token.

The text segmentation and syntactic analysis steps are 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.

1.2 Concepts Extraction

Concepts extraction is a process of discovering potential candidates for concepts in a CM. Generally, the subject of a sentence represents the first concept, and the second concept is represented by the object of a sentence [41].

The extraction of concepts is based on the textRank for keyword extraction algorithm [80] which is a variation of PageRank [81]. TextRank for keyword extraction algorithm allows to obtain the most important keywords in a document without the need of a training corpus or labeling and allows the use of the algorithm with different languages.

TextRank transfers the document into a graph of words, in which an edge between words stands for a relation between words. The importance of a word is determined by the importance of its neighbours. Formally, G = (V,E) denotes a directed graph with the set of vertices V and set of edges E. For a given vertex Vi, In(Vi) denote the set of vertices with edges to Vi, Out(Vi) the set of vertices with edges from Vi. The score of a vertex Vi is defined as follows:

$$S(Vi) = (1 - d) + d * \sum_{j \in In(Vi)} \frac{1}{|Out(Vi)|} S(Vj) \qquad (1)$$

where d is a damping factor that can be set between 0 and 1, which has the role of integrating into the model the probability of jumping from a given vertex to another random vertex in the graph usually, is set to 0.85, and this is the value we are also using in our implementation.

At the end of this phase, once a final score is obtained for each word in the graph, words are sorted in reversed order of their score, and N words with the highest score are chosen as a set of concept candidates. While N may be set to any fixed value.

1.3 Creation of Propositions

In this stage, we search for relationships between concepts. The process of finding a relationship in key sentences is based on the determination of syntagmatic[3] relations between words as triple "subject-predicate-object". For this purpose we propose the Triplet Extraction algorithm which is an extension of triplet algorithm described in [82].

[3] At the lexical level, syntagmatic structure in a language is the combination of words according to the rules of syntax for that language. For example, English uses noun + determiner + adjective.

The key sentence is every sentence in which at least two concepts from the set of concept candidates occur.

We start with parsing a sentence by Stanford parser and storing the result so that it can be taken as input for the proposed triplet extraction algorithm.

To find the multiple subjects in a sentence the algorithm searches the NP sub tree. The predicate is found in the VP sub tree and the objects are found in three different sub trees, all siblings of the VP sub tree containing the predicate. The sub trees are: PP (prepositional phrase), NP (noun phrase) and ADJP (adjective phrase).

As shown in Figure 4.2 the Triplet Extraction Algorithm takes as input the parte-of-speech (POS) of each word, the parse tree and the typed dependencies. Two functions are then called, the first is the GET_TRIPLETS and the second is the GET_RELATIONSHIP.

Function PREPROCESSOR (*sentence*)
Returns POS tagging, Parse tree, Typed Dependencies
// Run the Stanford parser with sentence as input
Output_sent ← i) POS of each word // see Table 4.2
ii) The parse tree generated // see Figure 4.6
iii) The typed dependencies // see Figure 4.7
Return *Ourput_sent*

Function TRIPLET_EXTACTION (*Output_sent*)
Return a solution, or error message
Function GET_TRIPLETS (*Output_sent*)
Function GET_RELATIONSHIP (*Output_sent*)

Figure 4.2 The Triplet Extraction Algorithm.

As shown in Figure 4.3 the GET_TRIPLETS function takes as input the Stanford Parse Tree and by considering the nodes under the NP sub tree and the VP sub tree, finds all the subjects, objects and predicates.

```
Function GET_TRIPLETS (Output_sent)
Return Multiple subjects, objects and predicates
if tree contains 'NP' then
  Function GET_SUBJECT (NP_subtree)
else
  Return error message
if tree contains 'VP' then
  Function GET_PREDICATE (VP_subtree)
  Function GET_OBJECT (VP_subtree)
else
  Return error message

Function GET_SUBJECT (NP_subtree)
Return Subject(s) and adjective(s)
for (all nodes of NP_subtree) do
  if NP_subtree contains 'NN?' then
    Store POS as a subject
    if NP_subtree contains 'JJ?' then
      Store POS as an adjective
Return the subject(s) and adjective(s)

Function GET_PREDICATE (VP_subtree)
Return Predicate(s)
for (all nodes of VP_subtree) do
  if VP_subtree contains 'VB?' then
    Store POS as a predicate
  else
    Return error message
Return the predicate(s)

Function GET_OBJECT (VP_subtree)
Return Object(s)
for (all nodes of VP_subtree) do
  if VP_subtree contains 'NP' then
    for (all nodes of VP_NP_subtree) do
      if VP_NP_subtree contains 'NN?' then
        Store POS as an object
      else
        Return error message
  else
    Return error message
Return the object(s)
```

Figure 4.3 The GET_TRIPLETS function.

The GET_RELATIONSHIP function finds the relationships between the subjects and objects. The algorithm is displayed in Figure 4.4.

```
Function GET_RELATIONSHIP (Output_sent)
Return relations
// Read the Stanford typed dependencies from Output_sent
for (all terms in typed_Dependencies) do
  if typed_Dependencies contain 'nsubj' then
    Store both words of nsubj as S1 and S2
    for each value of subject from GET_SUBJECT do
      if subject matches S2 then
      // Check for predicates
        for each value of predicate from GET_PREDICATE do
        if predicate matches S1 then
          Store S1 and S2  in the relation
  if typed_Dependencies contain 'dobj' or 'prep' then
    Store both words of dobj or prep as D1 and D2
    for each value of object in GET_OBJECT do
      if object matches D2 then
        Store value of object D2 in the relation
Store relation in relations
Return relations
```

Figure 4.4 The GET_RELATIONSHIP Function.

1.4 Ranking

The result of the ranking phase is the CM that provides an overview of the document's content.

In the first stage, the text ranking for keyword extraction algorithm is used for extracting the candidate concepts based on their score, this candidate concepts are used in the next stage for the creation of propositions. The importance of propositions is calculated based on the score of the concepts. Propositions with the higher calculated values are positioned higher in the CM hierarchy and the strongest proposition among of his calculated values is marked as the starting one.

1.5 Concept Map Visualisation

In order to complete the CMM process, the propositions need to be visualised as a concept map. CmapTools is concept mapping software developed by the Florida Institute for Human and Machine Cognition (IHMC) capable of importing triple written using 'CMap outline', 'CXL', 'XTM/XCM' or 'IVML' format [79]. The Concept Map Mining from text converted the propositions into CXL (Concept Map Extensible Language) format, an XML-based, light-weight file format to store concept map. Table 4.1 illustrates the most commonly used elements and attributes of CXL and Figure 4.5 shows the structure of an example CXL file.

CXL element	Attributes	Parent	Description
cmap	None	None	Main element
map	Root-id Width Height	cmap	Defines the structure of the map
concept-list	None	cmap	List of concepts in the map
linking-phrase-list	None	cmap	List of linking phrases in the map
connection-list	None	cmap	List of connections in the map
concept	id label	concept-list	Defines the concept
linking-phrase	id label	linking-phrase-list	Defines the linking phrase
connection	id from-id to-id	connection-list	Defines the connection

Table 4.1 Elements and attributes of CXL.

The generated CXL file can be directly imported to CMapTools using the option 'File -> import -> CMap from CXL file'. CMapTools provides a simple, user friendly interface for concept mapping, including auto layout, editing, sharing in the web, attaching related resources and merging with other concept maps [79].

```
<?xml version="1.0" encoding="UTF-8"?>
<cmap xmlns="http://cmap.ihmc.us/xml/cmap/"
      xmlns:dc="http://purl.org/dc/elements/1.1/">

    <map width="426" height="176">
        <concept-list>
            <concept id="1" label="Plants"/>
            <concept id="2" label="Leaves"/>
        </concept-list>
        <linking-phrase-list>
            <linking-phrase id="3" label="have"/>
        </linking-phrase-list>
        <connection-list>
            <connection id="XXX" from-id="1" to-id="3"/>
            <connection id="XXXX" from-id="3" to-id="2"/>
        </connection-list>
    </map>
</cmap>
```

Figure 4.5 Sample CXL file.

2 Natural Language Annotation

This section explains the NLP annotations: part-of-speech tagging, lemmatization and Stanford parser that we used in our method.

2.1 Part-Of-Speech Tagging

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective [83]. Table 4.2 summarises most commonly used part-of-speech tags.

Tag	description	Example
CC	Coordinating conjunction	and, but, or
DT	Determiner	a, the
IN	Preposition or subordinating conjunction	of, in, by
JJ	Adjective	yellow
NN	Noun, sing. or mass	llama
NNS	Noun (plural)	llamas
NNP	Proper noun, sing.	IBM
NNPS	Proper noun (plural)	Carolinas
PRP	Personal pronoun	I, you, he
PRP$	Possessive pronoun	Your, one's
RB	Adverb	quickly, never
RP	Particle	up, off
VB	Verb base form	eat
VBD	Verb past tense	ate
VBG	Verb gerund	eating
VBN	verb past participle	eaten
VBP	verb non-3sg pres	eat
VBZ	verb 3sg pres	eats
.	Sentence-final punctuation	. ! ?

Table 4.2 Part-of-speech tags [83].

2.2 The Stanford Parser

The Stanford Parser is a probabilistic parser which uses the knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences. This package is a Java implementation of probabilistic natural language parsers.

2.2.1 The Stanford Dependencies

The Stanford dependencies provide a representation of grammatical relations between words in a sentence for any user who wants to extract textual relationships. The dependency obtained from Stanford parser can be mapped directly to graphical representation in which words in a sentence are nodes in graph and grammatical relationships are edge labels. Figure 4.6 illustrates the Stanford Dependencies for the sentence '*Bell, based in Los Angeles, makes and distributes electronic, computer and building products*' [84].

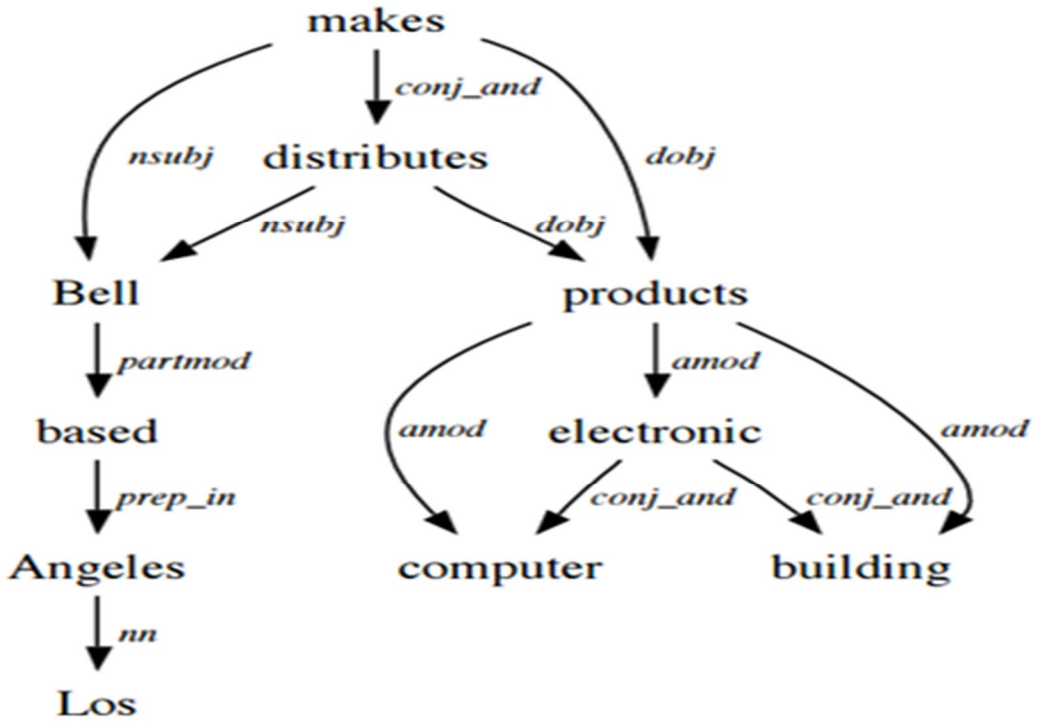

Figure 4.6 Graphical representation of the Stanford Dependencies for the sentence 'Bell, based in Los Angeles, makes and distributes electronic, computer and building products'[84].

2.2.2 The Parse Tree

The parse tree generated by the Stanford Parser is represented by three divisions: A sentence (S) having a noun phrase (NP), a verbal phrase (VP) and the full stop (.). The root of the tree is S. Figure 4.7 illustrates the parse tree for the sentence '*the cat set under the chair*'.

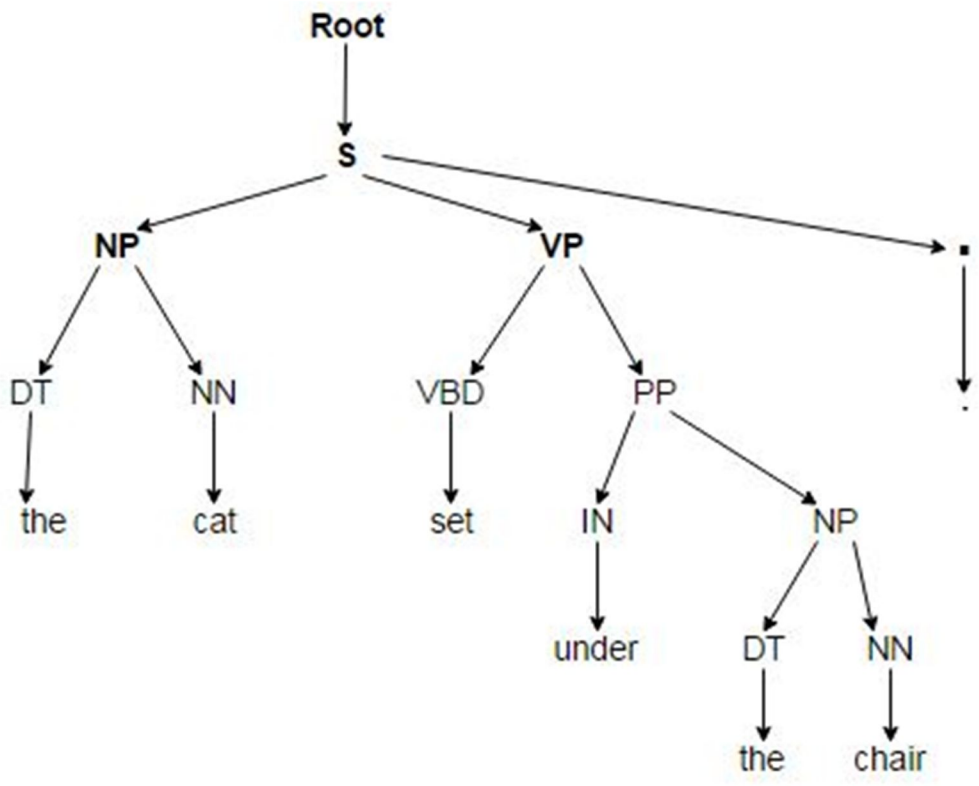

Figure 4.7 Graphical representation of the parse tree for the sentence 'the cat set under the chair'.

CONLUSION

An automatic generation of a human-quality CM from a given document is a difficult task. However, the research in this area has shown promising results. Our work is a contribution to the generation of CM that employs a new method that uses an unsupervised algorithm and NLP tools.

In this thesis, we presented the notion of concept maps, their uses and how they are constructed in chapter 2. Chapter 3 presented an overview of CMM process and discussed some of the existing approaches related to CMM. In chapter 4 a method for generating CMs from text was proposed and described in detail.

Our method for concept map generation consists of five stages: (1) preprocessing: this stage is responsible for providing all the resources needed for performing the extraction process; (2) concepts extraction: this stage involves the extraction of potential candidates for concepts in a CM using our triplet extraction algorithm; (3) creation of propositions: in this stage, we search for relationships between concepts; (4) ranking: in this stage, the propositions are ranked based on their concepts importance; (5) visualization: the generated concept map is visualised using IHMC CmapTools [79] by translating them into the concept map extensible language (CXL).

In our future work, we plan to experiment with other ranking and scoring strategies for concept extraction, enhance our method by including the anaphora resolution process and adapt our method to text written in Arabic.

References

[1] Novak, J. D. and A. J. Cañas., The Theory Underlying Concept Maps and How to Construct Them. Pensacola FI: Florida Institute for Human and Machine cognition, 2006.

[2] Canas, A, J., and Novak, J. D., Constructing your First Concept Map, 2009, from http://cmap.ihmc.us/docs/ConstructingAConceptMap.html .

[3] Novak, J. D., and Gowin, D. B. Learning How to Learn., New York,Cambridge University, 1984 .

[4] Gouli E, Gogoulou A, Papanikolaou K and Grigoriadou M. COMPASS, An adaptive web-based concept map assessment tool. Proceedings of the first international conference on concept mapping, Pamplona, 2004, pp. 295–302.

[5] Chang S.N. Externalising students' mental models through concept maps. Journal of Biological Education, 41, 2007, pp. 107–112.

[6] Van Boxtel C, Van der Linden J and Kanselaar G., Collaborative learning tasks and the elaboration of conceptual knowledge. Learning and Instruction; 10, 2000, pp. 311–330.

[7] Kinchin I.M, Hay D.B and Adams A., How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. Educational Research; 42, 2000, pp. 43–57.

[8] Willerman M. and MacHarg R.A., The concept map as an advance organizer. Journal of Research in Science Teaching; 28, 2006, pp. 705–711.

[9] Okebukola P.A., Can good concept mappers be good problem solvers in science? Research in Science and Technological Education; 10, 1992, pp. 153–170.

[10] Lloyd C.V., The elaboration of concepts in three biology textbooks: Facilitating student learning. Journal of Research in Science Teaching; 27, 1990, pp. 1019–1032.

[11] Zouaq A and Nkambou R., Building domain ontologies from text for educational purposes. IEEE Transactions on Learning Technologies**,** 27, 2008, pp. 49–62.

[12] Anderson, J. R., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y., An integrated theory of the mind. Psychological Review, 111, 2004, pp. 1036–1050.

[13] Markham, K. M., Mintzes, J. J., & Jones, M. G., The concept map as a research and evaluation tool: Further evidence of validity. Journal of Research in Science Teaching, 31, 1994, pp. 91-101.

[14] Alves, A., Pereira, F., and Cardoso, A., Automatic reading and learning from text. In Proceedings of the International Symposium on Artificial Intelligence (ISAI'2001), 2001, pp. 302-310.

[15] Novak, J. D. and Canas, A. J., The origins of the concept mapping tool and the continuing evolution of the tool. Information Visualization, 5, 2006, pp. 175-184.

[16] Ausubel, D. P., Novak, J. D., & Hanesian, H., Educational Psychology: A Cognitive View (2nd ed.). New York: Holt, Rinehart and Winston, 1978.

[17] Novak, J. D. and Gowin, D. B., Learning how to learn. Cambridge University Pres, Cambridge, United Kingdom, 1984.

[18] Sijtsma, B., Semi-Automatic Construction of Skeleton Concept Maps from Case Judgement Documents, Bachelor thesis, University of Amsterdam, 2014.

[19] Shen, R., Richardson, R, and Edward A. F., Concept maps as visual interfaces to digital libraries: summarization, collaboration, and automatic generation, 2003.

[20] Harith A., Sanghee K., David E. Millard, Mark J. Weal, Wendy H., Paul H. Lewis, and Nigel R., Automatic ontology-based knowledge extraction from web documents. IEEE Intelligent Systems, 18, 2003, pp. 14-21.

[21] Herl, H. E. Baker, E. L. and Niemi, D., Construct validation of an approach to modeling cognitive structure of U.S history knowledge. In Journal of Educational Research, 89, 1996, pp. 206-218.

[22] Markham, K. M. Mintzes, J. J. and Jones, M. G., The concept map as a research and evaluation tool: Further evidence of validity. Journal of Research in Science Teaching, 31, 1994, pp. 91-101.

[23] Richardson R. and Edward A. Fox Using concept maps in digital libraries as a cross-language resource discovery tool, In JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, Denver, CO, USA, 2005, pp. 256-257.

[24] Darmofal, D.L. Soderholm, D.H. and Brodeur, D.R., Using concept maps and concept questions to enhance conceptual understanding, Annual Conference, Montreal, Canada, 2002.

[25] Ryen W., Hyunyoung S., and Jay L., Concept maps to support oral history search and use, Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 2006, pp. 192-193.

[26] Maria, J., Concept mapping and appropriate instructional strategies in promoting programming skills of holistic learners, annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology, , Republic of South Africa, 2003, pp 308-315.

[27] Rendas, A.B. Fonseca, M. and Pinto, P.R., Toward meaningful learning in undergraduate medical education using concept maps in a PBL pathophysiology course. Adv Physiol Educ, 30, 2006, pp. 23-9.

[28] McClure, J.R. Sonak, B. and Suen H.K., Concept map assessment of classroom learning: Reliability, validity, and logical practicality. Journal of Research in Science Teaching, 1999, pp. 475-492.

[29] Johnsen, J.A. Biegel, D.E. and Shafran, R., Concept mapping in mental health: uses and adaptations. Evaluation and Program Planning, 23, 2000, pp. 67-75.

[30] Robert R. Hofman, John W. Coffey, Mary Jo Carnot, and Joseph D. Novak, An empirical comparison of methods for eliciting and modeling expert knowledge. In Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting. Santa Monica, 2002, pp. 482-486.

[31] Coffey, J.W. Canas, A.J. Hill, G. Carff, R. Reichherzer, T. and Suri, N., Knowledge modeling and the creation of el-tech: a performance support and training system for electronic technicians. Expert Systems with Applications, 25, 2003, pp. 483-492.

[32] Diana C.R., Joseph M.R., and Sara M.S., Using concept maps to assess student learning in the science classroom: Must different methods compete? , Journal of Research in Science Teaching, 35, 1998, pp. 1103-1127.

[33] da Rocha, F.E.L. Costa Jr, and Favero, E.L., A new approach to meaningful learning assessment using concept maps: ontologies and genetic algorithms In

Proceedings of the First International Conference on Concept Mapping ,Pamplona, Spain, 2004.

[34] Funaoi, H., Yamaguchi, E., and Inagaki, S., Collaborative concept mapping software to reconstruct learning processes. In Proceedings of the International Conference on Computers in Education, Washington, DC, USA, 2002, pp. 306.

[35] David B. Leake, Ana Maguitman, Thomas Reichherzer, Alberto J. Canas, Marco Carvalho, Marco Arguedas, So_a Brenes, and Tom Eskridge, Aiding knowledge capture by searching for extensions of knowledge models. Proceedings of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA, 2003, pp. 44-53.

[36] Pereira, F.C. Oliveira, A. and Cardoso, A., Extracting concept maps with clouds, In Proceedings of the Argentine Symposium of Artificial Intelligence (ASAI), Coimbra, Portugal, 2000.

[37] Andrew E.S and Michael S.H, Evaluation of unsupervised semantic mapping of natural language with leximancer concept mapping, Behaviour Research Method, 38, 2005, pp. 262-279.

[38] Brian R.G and Mildred L.G., Using knowledge acquisition and representation tools to support scientific communities. In Proceedings of the twelfth national conference on Artificial intelligence, 1 , 1994, pp. 707-714,

[39] Bruillard, E. and Baron, G.L., Computer-based concept mapping: a review of a cognitive tool for students. In Proceedings of Conference on Educational Uses of Information and Communication Technologies, Publishing House of Electronics Industry (PHEI), 2000, pp. 331-338.

[40] Villalon, J. J..and Calvo, R. A., Concept map mining: A definition and a framework for its evaluation, in Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 3, 2008, pp. 357–360.

[41] Zubrinic, K., Kalpic, D., and Milicevic, M., The automatic creation of concept maps from documents written using morphologically rich languages. Expert Systems with Applications, 39, 2012, pp. 12709–12718.

[42] Trochim, W. M. K., An introduction to concept mapping for planning and evaluation, Evaluation and Program Planning, 12, 1989, pp. 1–16.

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

[44] Das, D. and Martins, A. F. T., A survey on automatic text summarization, Language Technologies Institute, 4, 2007, pp 1-31.

[45] Spärck K.J., Automatic summarising: The state of the art, Information Processing and Management, 43, 2007, pp. 1449–1481.

[46] Binwahlan, M. S. Salim, N. and Suanmali, L., Fuzzy swarm based text summarization, Journal of Computer Science, 5, 2009, pp. 338–346.

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

[48] Willis, C. L. and Miertschin, S. L., Centering resonance analysis: a potential tool for IT program assessment, in Proceedings of the 2010 ACM conference on

Information technology education, ser. SIGITE '10. New York, NY, USA: ACM, 2010, pp.135-142.

[49] chen, N.S. Kinshuk,. Wei, C.W and Chen, H.-J., Mining e-learning domain concept map from academic articles, Computers & Education, 50, 2008, pp. 1009-1021.

[50] Furdik, K., Paralic J. and smrz P., Classification and automatic concept map creation in eLearning environment, In Proceedings of the Czech-Slovak scientific conference Znalosti, 2008.

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

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

[53] Oliveira, A. Pereira, F.C. and Cardoso, A., Automatic reading and learning from text, In Proceedings of the International Symp. On Artificial Intelligence, 2002, pp. 1–12.

[54] Kof, L. Gacitua, R. Rouncefield, M. and Sawyer, P., Concept mapping as a means of requirements tracing, In Managing Requirements Knowledge (MARK), Sydney, Australia, 2010.

[55] Rajaraman K. and Tan, A.-H., Knowledge discovery from texts: a concept frame graph approach, In Proceedings of the 11th International Conference on Information and Knowledge Management, New York, NY, USA, 2002.

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

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

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

[59] Sue, P.-C. Weng, J.-F. Su, J.-M and Tseng, S.-S., A new approach for constructing the concept map, IEEE International Conference on Advanced Learning Technologies, 49, 2004, pp. 76–80.

[60] Cañas A. J., Carvalho, M. Arguedas M., Leake, D. B Maguitman A. and Reichherzer, T., Mining the web to suggest concepts during concept map construction, In Proceedings of the First International Conference on Concept Mapping, Pamplona, Spain, 2004.

[61] Atapattu, M., Tushari, D.A., A computational model for task-adapted knowledge organisation: improving learning through concept maps extracted from lecture slides., Thesis (Ph.D.), University of Adelaide, School of Computer Science, 2015.

[62] Valerio, A., & Leake, D. B., Using Automatically Generated Concept Maps for Document Understanding: A Human Subjects Experiment, Paper presented at the Fifth International Conference on Concept Mapping, Valletta, Malta, 2012.

[63] Olney, A. M., Cade, W. L., and Williams, C., Generating concept map exercises from textbooks. Paper presented at the Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications, Portland, Oregon, 2011.

[64] Richardson, R., Using concept maps as a tool for cross-language relevance determination, PhD thesis, Blacksburg, Virginia, 2007.

[65] Villalon, J. and Calvo, R. A., Concept extraction from student essays, towards concept map mining, In Proceedings of the Ninth IEEE International Conference on Advanced Learning Technologies, Washington, DC, USA, 2009.

[66] Wang, W., Cheung, C., Lee, W. and Kwok, S., Mining knowledge from natural language texts using fuzzy associated concept mapping, Information Processing and Management, 44, 2008, pp. 1707–1719.

[67] Chen, S.-M. & Bai, S.-M., Using data mining techniques to automatically construct concept maps for adaptive learning systems, Expert Systems with Applications, 37, 2010, pp. 4496–4503.

[68] Lau, R. Y. K., Chung, A. Y. K., Song, D., and Huang, Q., Towards Fuzzy Domain Ontology Based Concept Map Generation for E-Learning, Springer Berlin Heidelberg, 4823, 2008, pp. 90-101.

[69] Lau, R. Y. K., Song, D., Li, Y., Cheung, T. C. H. & Hao, J.-X., Toward a fuzzy domain ontology extraction method for adaptive e-learning, IEEE Transactions on Knowledge and Data Engineering, 21, 2009, pp. 800–813.

[70] Lee, C.-H., Lee, G.-G. & Leu, Y., Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning, Expert Systems with Applications, 36, 2009, pp. 1675–1684.

[71] Villalon, J., & Calvo, R. A., Concept maps as cognitive visualisations of writing assignments. Educational technology and Society, 14, 2011, pp. 16-27.

[72] Miller, G. A., WordNet: a lexical database for English, Communications of the ACM, 38, 1995, pp. 39–41.

[73] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V., GATE: A framework and graphical development environment for robust NLP tools and applications. Paper presented at the Proceedings of the 40th Annual Meeting of the ACL, Philadelphia, 2002.

[74] Fisher, K. M., Wandersee, J. H., and Moody, D. E., Mapping Biology Knowledge, Springer Netherlands, 2002.

[75] Manning, C., Raghavan, P., & Schutze, H., Introduction to Information Retrieval, Cambridge University Press New York, NY, USA, 2008,

[76] Gruber, T. R., A translation approach to portable ontology specifications, Knowledge Acquisition, 5, 1993, pp. 199–220.

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

[78] Toutanova, K., Klein, D., Manning, C., & Singer, Y., Feature-rich part-of speech tagging with a cyclic dependency network, Paper presented at the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology Edmonton, Canada, 2003.

[79] Canas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., and Eskridge, T., CmapTools: A knowledge modeling and sharing environment, Paper presented at the First International Conference on Concept Mapping, Pamplona, Spain, 2004

[80] Mihalcea R. and Tarau. P., TextRank: Bringing Order into Texts, In Conference on Empirical Methods in Natural Language Processing, 2004, pp. 404–411.

[81] Page, L., Brin, S., Motwani, R., Winograd, T., The pagerank citation ranking: Bringing order to the web, In Proceedings of the 7th International World Wide Web Conference., Brisbane, Australia, 1998, pp. 161–172.

[82] Rusu, D., Dali, L., Fortuna, B., Marko, G., & Dunja, M., Triplet Extraction from Sentences. Paper presented at the Data mining and Data warehouses, Ljubljana, Slovenia, 2007.

[83] Jurafsky, D., Martin, J. H., Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition. Prentice Hall, Pearson Education International, 2009, ISBN: 9780135041963.

[84] de Marneffe, M.-C. and Manning, C. D., The Stanford typed dependencies manual, in Revised for Stanford Parser v. 3.7.0 2016, 2008.

ملخص

تعرف الخرائط المفاهيمية بأنها أداة تخطيطية لعرض مجموعة من المفاهيم ضمن شبكة من العلاقات بحيث يتم ترتيب المفاهيم بشكل هرمي من الأكثر عمومية وشمولية إلى الأقل عمومية، ويتم الربط بين المفاهيم بخطوط يكتب عليها جملة أو كلمة ذات معنى علمي تسمى (الكلمات الرابطة). تعتبر الخرائط المفاهمية أداة قوية تستخدم في التعليم و التعلم، فهي تساعد الطلبة على الابداع والتفكير التأملي و حل المشكلات. يتم بناء الخرائط المفاهمية عادة إما يدويا أو تلقائيا. يستخدم مصطلح " Concept Map Mining" للإشارة إلى البناء التلقائي للخرائط المفاهمية من وثيقة نصية. هذه الأطروحة تقدم لمحة عامة عن عملية البناء التلقائي للخرائط المفاهمية و الطرق المختلفة المستعملة في ذلك، كما أننا قمنا بتقديم طريقة جديدة لعملية البناء التلقائي للخريطة المفاهمية من نص.

الكلمات المفتاحية: الخرائط المفاهمية، المفاهيم، الكلمات الرابطة، البناء التلقائي للخرائط المفاهمية

Abstract

Concept maps are graphical tools for organizing and representing knowledge. Many educators and researchers have exploited CMs in a number of ways, including evaluation or assessment tools, knowledge management and cooperative meaningful learning tools in education. The construction of concept maps is typically done either manually or automatically. The term Concept Map Mining (CMM) is used to refer to the automatic generation of Concept Map from document. This thesis gives an overview of concept map Mining and their different approaches, then we introduce our method for automatic generation of concept map from a text. The proposed method uses an unsupervised algorithm and NLP tools.

Keywords: Concept maps, Concept Map Mining (CMM), unsupervised algorithm, NLP tools

Résumé

Un schéma conceptuel (ou carte conceptuelle) est un outil qui permet d'organiser et de représenter graphiquement la structure des connaissances. Le schéma conceptuel poursuit plusieurs buts. Il permet de représenter le modèle mental d'une situation, que cette représentation soit personnelle, celle d'un groupe ou d'une organisation. Il permet aussi de résumer la structure de la connaissance extraite d'ouvrages écrits. La construction de carte conceptuelle est généralement effectuée manuellement ou automatiquement. Le terme « Concept Map Mining » permet de se référer à la construction automatique de la carte conceptuelle à partir du document. Ce mémoire donne un aperçu général de la construction automatique des cartes conceptuelles et de leurs différentes approches. Ensuite, nous présentons notre méthode pour la construction automatique de la carte conceptuelle à partir d'un texte. La méthode proposée utilise un algorithme non supervisé et des outils NLP.

Mots-clefs: carte conceptuelle, construction automatique des cartes conceptuelles, algorithme non supervisé, des outils NLP