UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Camila Zacché de Aguiar

Concept Maps Mining for Text Summarization

VITÓRIA-ES, BRAZIL
March 2017

Camila Zacché de Aguiar

Concept Maps Mining for Text Summarization

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do Grau de Mestre em Informática.

Orientador (a): Davidson Cury
Co-orientador: Amal Zouaq

VITÓRIA-ES, BRAZIL
March 2017

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Setorial Tecnológica,
Universidade Federal do Espírito Santo, ES, Brasil)

Aguiar, Camila Zacché de, 1987-
A282c Concept maps mining for text summarization / Camila Zacché de Aguiar. – 2017.
149 f. : il.

Orientador: Davidson Cury.
Coorientador: Amal Zouaq.
Dissertação (Mestrado em Informática) – Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Informática na educação. 2. Processamento de linguagem natural (Computação). 3. Recuperação da informação. 4. Mapas conceituais. 5. Sumarização de Textos. 6. Mineração de dados (Computação). I. Cury, Davidson. II. Zouaq, Amal. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 004

Mestranda Camila Zacche de Aguiar

Concept Maps Mining for Text Summarization

Camila Zacche de Aguiar

Dissertação submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do grau de Mestre em Informática.

Aprovada em 31 de março de 2017:

Prof. Dr. Davidson Cury (PPGI - UFES)

Profª. Drª. Amal Zouaq (UOTTAWA-CANADÁ)

Elias de Oliveira

Prof. Dr. Elias Silva de Oliveira (PPGI-UFES)

Aline Villavic

Profª. Drª. Aline Villavicencio (UFRGS)

Prof. Dr. Crediné Silva de Menezes (PPGI-UFES)

“Most of the fundamental ideas of science are essentially simple, and may, as a rule, be expressed in a language comprehensible to everyone.”

Albert Einstein

Acknowledgments

I would like to thank the many people who have been with me over the years and who have contributed in one way or another to the completion of this master's degree.

Especially my advisor, *Davidson Cury*, who in a constructivist way disoriented me several times as a stimulus to the search for new answers. Thanks for your friendship, availability, reliability and so many shares of knowledge and challenge. Mainly for opening the doors of the academic world to me and supporting my research interests.

I also want to thank *Amal Zouaq*, who has been a wonderful co-advisor, for having accepted to guide me with so much availability and attention, to have shared knowledge and experiences, to have believed in my work and promptly having extended her hand.

To the professors of the Federal University of Espírito Santo (UFES): Prof. *Tania Gava* for her contributions and ideas; to Prof. *Crediné Menezes* for his rich remarks and questions; to Prof. *Orivaldo Tavares* for his attention and Prof. *Victor Souza* for all learning and support.

To master's friends, by the exchange of knowledge and experiences. Mainly to *Wagner* for introducing me the concept maps; *Marcos* for the long hours of discussion and learning; and *Patricia* for all shared experiences.

The faculty members of the Examining Committee for having heeded the invitation to play this role, offering their time and expertise.

To CAPES (Commission for the Improvement of Higher Education Personnel) by providing better conditions for the development of this work.

To my family and friends who always supported me. To my parents, *Ilta* and *Lenine*, for the encouragement, support and love. To my brothers *Rodolfo* and *Frederico* for all assistance, patience and knowledge. To *Marcelo* for all love and understanding.

To *God*, for allowing all this to be possible.

Abstract

Concept maps are graphical tools for the representation and construction of knowledge. Concepts and relationships form the basis for learning and, therefore, concept maps have been extensively used in different situations and for different purposes in education, one of them being representation of written text. Even a complex and grammatically difficult one can be represented by a concept map containing only concepts and relationships that represent what was expressed in a more complicated way.

However, the manual construction of a concept map requires quite a bit of time and effort in the identification and structuring of knowledge, especially when the map should not represent the concepts of the author's cognitive structure. Instead, the map should represent the concepts expressed in a text. Thus, several technological approaches have been proposed in order to facilitate the process of constructing concept maps from texts.

This dissertation proposes a new approach to automatically build concept maps as a summarization of scientific texts. The summarization aims to produce a concept map as a summarized representation of the text while maintaining its various and most important characteristics.

The summarization facilitates the understanding of texts, as the students are trying to cope with the cognitive overload caused by the increasing amount of available textual information. This increase can also be harmful to the construction of knowledge. Thus, we hypothesized that the summarization of a text represented by a concept map may contribute for assimilating the knowledge of the text, as well as decrease its complexity and the time needed to process it.

In this context, we conducted a review of literature from between the years of 1994 and 2016 on the approaches aimed at the automatic construction of concept maps from texts. From it, we built a categorization to better identify and analyze the features and characteristics of these technological approaches. Furthermore, we sought to identify the limitations and gather the best features of the related works to propose our approach.

Besides, we present a process for Concept Map Mining elaborated following four dimensions: Data Source Description, Domain Definition, Elements Identification and Map Visualization.

In order to develop a computational architecture to automatically build concept maps as summarization of academic texts, this research resulted in the public tool *CMBuilder*, an online tool for the automatic construction of concept maps from texts, as well as a public api java called *Extrout*NLP, which contains libraries for information extraction and public services.

In order to reach the proposed objective, we used methods from natural language processing and information retrieval. The main task to reach the objective is to extract propositions of the type (*concept, relation, concept*) from the text. Based on that, the research introduces a pipeline that comprises the following: grammar rules and depth-first search for the extraction of concepts and relations between them from text; preposition mapping, anaphora resolution, and exploitation of named entities for concept labeling; concepts ranking based on frequency and map topology; and summarization of propositions based on graph topology. Moreover, the approach also proposes the use of supervised learning techniques of clustering and classification associated with the use of a thesaurus for the definition of the text domain and the construction of a conceptual vocabulary of the domain.

Finally, an objective analysis to validate the accuracy of *Extrout*NLP library is performed and presents 0.65 precision on the corpus. Furthermore, a qualitative analysis to validate the quality of the concept map built by the *CMBuilder* tool is performed, reaching 0.75/0.45 for precision/recall of concepts and 0.57/0.23 for precision/recall of relationships in English language, and reaching 0.68/0.38 for precision/recall of concepts and 0.41/0.19 for precision/recall of relationships in Portuguese language. In addition, an experiment to verify if the concept map summarized by CMBuilder has influence for the understanding of the subject addressed in a text is conducted, reaching 60% of hits for maps extracted from small texts with multi-choice questions and 77% of hits for maps extracted from extensive texts with discursive questions.

Keywords: Concept Map, Concept Map Mining, Natural Language Processing, Information Retrieval, Summarization, Knowledge Representation.

Resumo

Os mapas conceituais são ferramentas gráficas para a representação e construção do conhecimento. Conceitos e relações formam a base para o aprendizado e, portanto, os mapas conceituais têm sido amplamente utilizados em diferentes situações e para diferentes propósitos na educação, sendo uma delas a representação do texto escrito. Mesmo um gramático e complexo texto pode ser representado por um mapa conceitual contendo apenas conceitos e relações que representem o que foi expresso de uma forma mais complicada.

No entanto, a construção manual de um mapa conceitual exige bastante tempo e esforço na identificação e estruturação do conhecimento, especialmente quando o mapa não deve representar os conceitos da estrutura cognitiva do autor. Em vez disso, o mapa deve representar os conceitos expressos em um texto. Assim, várias abordagens tecnológicas foram propostas para facilitar o processo de construção de mapas conceituais a partir de textos.

Portanto, esta dissertação propõe uma nova abordagem para a construção automática de mapas conceituais como sumarização de textos científicos. A sumarização pretende produzir um mapa conceitual como uma representação resumida do texto, mantendo suas diversas e mais importantes características.

A sumarização pode facilitar a compreensão dos textos, uma vez que os alunos estão tentando lidar com a sobrecarga cognitiva causada pela crescente quantidade de informação textual disponível atualmente. Este crescimento também pode ser prejudicial à construção do conhecimento. Assim, consideramos a hipótese de que a sumarização de um texto representado por um mapa conceitual pode atribuir características importantes para assimilar o conhecimento do texto, bem como diminuir a sua complexidade e o tempo necessário para processá-lo.

Neste contexto, realizamos uma revisão da literatura entre os anos de 1994 e 2016 sobre as abordagens que visam a construção automática de mapas conceituais a partir de textos. A partir disso, construímos uma categorização para melhor identificar e analisar os recursos e as características dessas abordagens tecnológicas. Além disso, buscamos identificar as limitações e reunir as melhores características dos trabalhos relacionados para propor nossa abordagem.

Ademais, apresentamos um processo Concept Map Mining elaborado seguindo quatro dimensões: Descrição da Fonte de Dados, Definição do Domínio, Identificação de Elementos e Visualização do Mapa.

Com o intuito de desenvolver uma arquitetura computacional para construir automaticamente mapas conceituais como sumarização de textos acadêmicos, esta pesquisa resultou na ferramenta pública *CMBuilder*, uma ferramenta online para a construção automática de mapas conceituais a partir de textos, bem como uma api java chamada *ExtroutNLP*, que contém bibliotecas para extração de informações e serviços públicos.

Para alcançar o objetivo proposto, direcionados esforços para áreas de processamento de linguagem natural e recuperação de informação. Ressaltamos que a principal tarefa para alcançar nosso objetivo é extrair do texto as proposições do tipo (*conceito, relação, conceito*). Sob essa premissa, a pesquisa introduz um pipeline que compreende: regras gramaticais e busca em profundidade para a extração de conceitos e relações a partir do texto; mapeamento de preposição, resolução de anáforas e exploração de entidades nomeadas para a rotulação de conceitos; ranking de conceitos baseado na análise de frequência de elementos e na topologia do mapa; e sumarização de proposição baseada na topologia do grafo. Além disso, a abordagem também propõe o uso de técnicas de aprendizagem supervisionada de clusterização e classificação associadas ao uso de um tesauro para a definição do domínio do texto e construção de um vocabulário conceitual de domínios.

Finalmente, uma análise objetiva para validar a exatidão da biblioteca *ExtroutNLP* é executada e apresenta 0.65 precision sobre o corpus. Além disso, uma análise subjetiva para validar a qualidade do mapa conceitual construído pela ferramenta *CMBuilder* é realizada, apresentando 0.75/0.45 para precision/recall de conceitos e 0.57/0.23 para precision/recall de relações em idioma inglês e apresentando 0.68/0.38 para precision/recall de conceitos e 0.41/0.19 para precision/recall de relações em idioma português. Ademais, um experimento para verificar se o mapa conceitual sumarizado pelo CMBuilder tem influência para a compreensão do assunto abordado em um texto é realizado, atingindo 60% de acertos para mapas extraídos de pequenos textos com questões de múltipla escolha e 77% de acertos para mapas extraídos de textos extensos com questões discursivas.

Palavras Chave: Mapa Conceitual, Mineração de Mapas Conceituais, Processamento de Linguagem Natural, Recuperação de Informação, Sumarização, Representação do Conhecimento.

List of Figures

List of Tables

Contents

Chapter 1
Introduction

This chapter presents an overview of the research conducted in the course of this work explaining ideas about context, motivation, hypotheses, questions, objectivies, methods, process, contributions and scientific productions. These explanations and discussions will guide all subsequent chapters.

1.1 Context

The information society is constantly accessing information very quickly and widely, and new information is produced, reflected, published or shared almost instantly. While this enables us to immerse ourselves in this vast information network, it also produces a cognitive overload.

The cognitive overload indicates that the perceptive and cognitive processes are overwhelmed by technological advances (TOFFLER, 1970), i.e., we are unable to absorb and process all the information to which we are exposed. According to the EMC Digital Universe study, it is estimated that 1 septillion bits of information were produced in the year 2014 and the expectation is that this number will be multiplied by 6 until the year 2020.

Analogously, academic data follow the same trend of growth. We conducted a quantitative analysis on all collections of some scientific databases (Springer, IEEE Xplore and ACM) to exemplify the growing amount of academic data published over the years. As shown in Figure 1.1, we observed that the number of articles published in the last sixteen years is higher than in the last century.

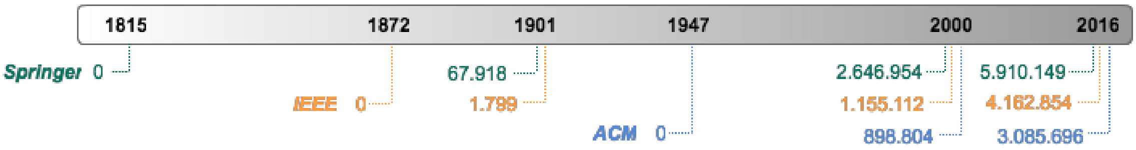


Figure 1.1 Number of articles published per period for the ACM, Springer and IEEE collections (by Author)

Therefore, a student is faced with a large amount of information in uncontrollable flow to keep informed about a particular subject. This brings different challenges into the student's learning process, from which we highlight:

(i) The student must select the relevant documents for a particular subject from among all available documents. Naturally, he/she needs to understand the information contained in a given document, i.e., he/she must invest time in reading the whole document in order to determine, whether to select it or not.

(ii) Documents are composed of a large amount of information usually written in complex form and, in most cases, the language used is different from the student's own language, which interferes with the student’s ability to read and understand the document in question.

(iii) After selecting the relevant document, the student must also invest a great deal of cognitive effort to identify and understand the information discovered.

Understanding this complexity and reflecting on the information requires considerable cognitive effort and time. In order to facilitate this process, concept maps can be used as a more meaningful representation of the information. According to Novak & Cañas (2010), concept maps are graphical tool for representing and organizing knowledge that are comprised of concepts and the relationships between them. Research from literature also suggests that graphical representation can reduce the problems of information overload and learning disorientation for learners (CHEN, et al., 2008). Therefore, a complex text can be represented by a concept map containing only concepts and relationships that summarize what was expressed in a more complex way.

Although the text representation by means of a concept map is an interesting resource for learning, its construction is still a challenge. The manual construction of a concept map requires a certain dedication of time and effort engaged in the identification and structuring of knowledge. Moreover, the construction of a concept map becomes more complex when the author does not representing his knowledge, but the knowledge expressed in a text written by another person.

In this context, we note the development of technological approaches that assist or automate the process of constructing concept maps from texts. These approaches adopt different techniques of Natural Language Processing (NLP) and Information Retrieval (IR). However, we can consider that the results are still not satisfactory and have some limitations that will be further discussed below. An automatically generated concept map, depicting the overview of a specific domain knowledge or documents, can facilitate the learner’s understanding of the content (LEE, et al., 2015).

The problem addressed in this research project can be defined by the following question: **How to automatically construct concept maps of scientific style for the summarization of academic texts?**

In research context, we adopted the term *text*, as a generalization of a written text, i.e., a graphic and visual representation of words sequences, depicted by letters, punctuation, diacritics and specific linguistic descriptions (PRETI, 2006); the term *summarization*, as a concise representation of the most important information contained in the text (SIDDHARTHAN, et al., 2011); and the term *scientific style*, as a concept map governed by two basic rules: the map might contain only concepts, and there is always a verb in a relationship between concepts (Section 4.1) (AGUIAR & CURY, 2016).

1.2 Motivation

Concept maps have been extensively used in education for different situations and purposes, such as a learning resource, means of evaluation, instructional organization, cognitive representation, elicitation and sharing of knowledge. In this scenario, maps can be used as tools to support education, since teachers use them to verify the student's level of understanding, to analyze the average knowledge of a class, to identify concepts and meanings wrongly assimilated or made explicit and shared knowledge about a study domain.

We also would like to emphasize the use of concept maps as a tool for the graphical representation of texts in order to provide a visual and holistic way to knowledge representation. Therefore, a more dynamic and flexible graphical representation composed of concepts and relationships is considered easier to be built, assimilated and understood than a complex text. Furthermore, the essential information of the text expressed by means of meaningful propositions allows a new viewpoint on the information. Thus, a single map can be interpreted in different ways depending on the reader and a single text can generate different maps depending on the author.

Using concept maps, the student could know the main concepts of the subject before plunging deeply into the text. This would favor the assimilation of new knowledge, especially in texts whose language is not the mother tongue of the student. Consequently, looking at the graphical representation of the concept map, the student could spend less time to analyze the relevance of the document to the subject.

Concept map representation of individual documents that effectively produces summaries of those documents allows users to get an understanding of the document without going through the entire document (KARANNAGODA, et al., 2013).

Since concept maps provide important benefits for learning, several approaches have emerged for the automatic construction of concept maps from texts. The first work in this context was the Gnosis system (GAINES & SHAW, 1994), whose goal was the use of knowledge acquisition techniques on the electronic documents used in the communication between the scientific communities involved in the project. In addition to this, we would like to point out approaches directed to maps construction as a part of a lightweight ontology (ZOUAQ, et al., 2007), as an index form (VALERIO, et al., 2008), and as a knowledge representation in learning virtual environments (LAU, et al., 2007), among others.

Although there are some approaches in this context, none of them is publicly available for use as a tool. Moreover, the approaches are directed neither to the texts summarization nor construction of concept maps in scientific style. Therefore, these are the main motivations that led to the development of this research.

1.3 Research Hypothesis

Based on the problems presented in Section 1.1, the following arguments were formulated as hypotheses:

(i) It is possible to create a tool for automatically building concept maps of scientific style to represent the summarization of a text.

(ii) The variation of linguistic components of a given conceptual model can provide multilingual application in Portuguese and English language.

(iii) The use of a domain knowledge base such as Thesaurus can improve the quality of the summarized concept map of a text.

(iv) The use of a concept map automatically summarized from a text influences the understanding of the subject addressed in that text.

1.4 Research Questions

The following formulates the main questions that this research aims to investigate:

(i) What web tool or service are available for the automatic construction of concept maps from texts?

(ii) What are the characteristics and limitations of the approaches proposed for the automatic construction of concept maps from texts in Portuguese and English languages?

(iii) What are the techniques and methods used to extract propositions from texts?

(iv) Is it possible to develop and adapt the NLP resources needed to extract propositions from text?

(v) What methods should be designed to assess whether the concept map constructed by the approach represents a summarization of a text?

1.5 Research Objectives

The general objective of this research is to **develop a computational architecture to automatically build concept maps of scientific style as summarization of academic texts**. The proposal is supported by several distinct techniques that complement each other in Natural Language Processing (NLP) and Information Retrieval (IR).

A secondary objective of this research is to study the influence that a concept map summarized from text has on the learning process. This kind of knowledge can then be used by researchers to develop new pedagogical strategies.

1.6 Research Methods

The research purpose is to find answers to questions by applying scientific methods (SELLTIZ, et al., 1967). Thus, we classify the research developed in this work as different sections namely Nature, Problem Approach, Objective, Procedures Technical, Scientific and Research. Such classification is synthesized in Figure 1.2 and explained below.

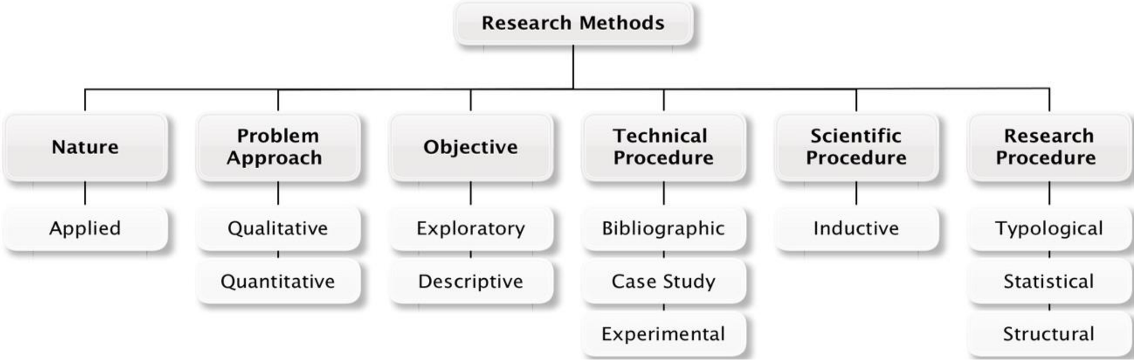


Figure 1.2 Synthesis of research method (by Author)

The **Nature**, following the classification proposed by Ander-Egg (ANDER-EGG, 1978), is defined as **Applied**, since the study has practical interest and the results are applied in solving the real problems such as a tool for automatically building concept maps for text summarization.

The **Problem Approach**, following the classification proposed by Sampieri, Collado & Lucio (SAMPIERI, et al., 2013), is defined as **Qualitative** for the development of the approach since the researcher was the key to understand and investigate phenomena from certain experiments and improve their results and as **Quantitative** for the objective analysis and comparison of results.

The **Objective**, following the classification proposed by Gil (GIL, 2008), is defined as **Exploratory** since the literature is reviewed in order to provide a subject overview. It is also characterized as **Descriptive** as it identifies and describes characteristics and features of technological approaches for the construction of concept maps from texts, besides using quizzes to collect data in the real context of manual construction of concept maps.

The **Technical Procedure**, following the classification proposed by Gil (GIL, 2008), is defined as **Bibliographic**, **Experimental** and **Case Study**, since we use theoretical references to collect information, apply variables for observation of effects during the development of the approach, and study the influence of concept maps summarized on the understanding of the text.

The **Scientific Procedure**, following the classification proposed by Marconi & Lakatos (MARCONI & LAKATOS, 2004) is defined as **Inductive** since we consider empirical knowledge and from experience to extract solutions. Thus, we start from concrete observations on the process and then generalize the solution into likely conclusions.

As for the **Research Procedure**, following the classification proposed by Marconi & Lakatos (MARCONI & LAKATOS, 2004), which targets a more practical view with a restricted purpose, it is defined as **Typological**, since we determine the characteristics of a new approach from the classification and comparison of similar approaches; **Statistical**, with the application of quantitative analysis on the experiments; and **Structural**, with the investigation of the concrete phenomenon on manual construction of concept maps, we come to an abstract level through of the conceptual architecture, and return to the concrete implementing the solution.

1.7 Research Process

The following presents the steps of the investigative process applied during this research. For this we follow the proposal of Quivy & Campenhoudt (QUIVY & CAMPENHOUDT, 2005), dealing with the investigative process as a theater play composed by three acts and seven scenes. Figure 1.3 synthesizes the process and highlights the results produced in each scene, which are considered as input for subsequent scenes.

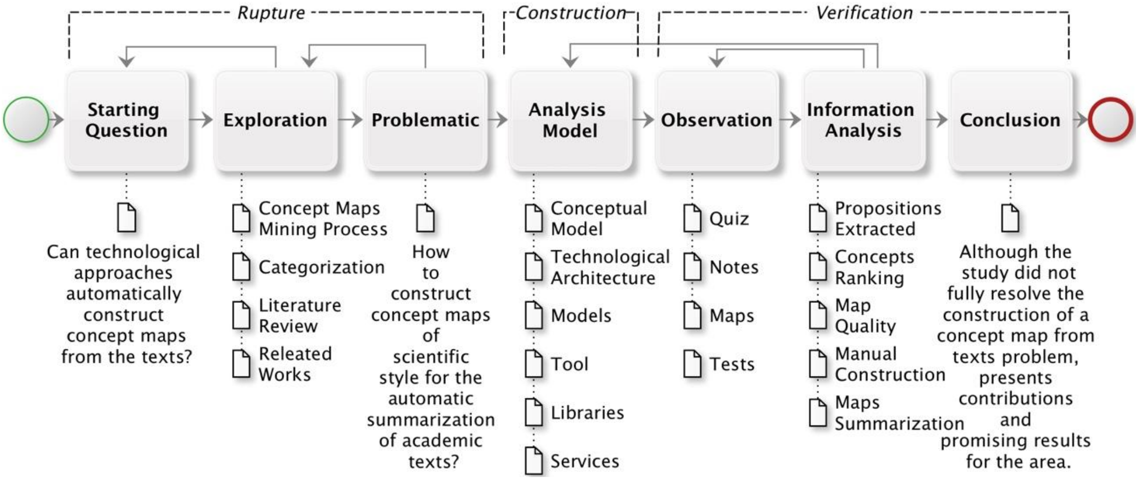

Rupture
Construction
Verification
Starting Question
Exploration
Problematic
Analysis Model
Observation
Information Analysis
Conclusion
Can technological approaches automatically construct concept maps from the texts?
Concept Maps Mining Process
Categorization
Literature Review
Releated Works
How to construct concept maps of scientific style for the automatic summarization of academic texts?
Conceptual Model
Technological Architecture
Models
Tool
Libraries
Services
Quiz
Notes
Maps
Tests
Propositions Extracted
Concepts Ranking
Map Quality
Manual Construction
Maps Summarization
Although the study did not fully resolve the construction of a concept map from texts problem, presents contributions and promising results for the area.

Figure 1.3 Synthesizes of research process (by Author)

The **Rupture** act breaks with the preconceptions and false evidence. This act resulted in a literature review on the technological approaches to construction of concept maps and led to the creation of a categorization representing the various methods and characteristics for automatic concept map mining from text. In this act we define the following scenes:

(i) **Starting Question:** Can technological approaches automatically construct concept maps from the texts?

(ii) **Exploration:** The exploration of subject was performed by the following methods:

- Literature review (Section 4.2) conducted by a systematic search in IEEEXplore Digital Library, ACM Digital Library, and Elsevier ScienceDirect. A total of 134 publications were collected, of which 55 were pre-selected, and 30 of which presented an approach relevant to our study.
- Investigation of public technological resources for NLP and IR tasks (APIs, services, etc.).
- Observation and analysis of the information cataloged from the categorization created on the technological approaches for the construction of concept maps from texts.

(iii) **Problem:** The problem was established based on the results obtained in the Exploration stage. We apply the categorization on similar approaches in order to identify the characteristics of each approach and its positive and negative points. Therefore, we define the problem "*How to construct concept maps of scientific style for the automatic summarization of academic texts?*" (Section 1.1).

The **Construction** act expresses the logic of the phenomenon studied and builds its propositions, research plan and operations. This act resulted in a conceptual model and technological architecture, library for information extraction tasks, models for Portuguese language, and a service and tool for the construction of concept maps from texts. In this act we define the following scene:

(i) **Analysis Model:** Developed from observations and experiments transformed in a systematic form. Based on the information collected in the rupture act, a conceptual model was defined and the necessary resources for the model implementation were developed and adapted, resulting in a service-oriented architecture. Finally, a tool was implemented following the proposed architecture.

The **Verification** act verifies the propositions by the facts. This act resulted in the analysis of results presented by the tool using precision, recall, comparison and questionnaire. In this act we define the following scenes:

(i) **Observation:** The observation was applied on components not strictly representative, but with characteristics of the population. Therefore, the *Direct Observation* was applied on the NLP tasks, allowing the identification of errors and improvements. *Indirect Observation* was applied by means of a quiz to obtain information about the quality and summarization of the map, besides of the difficulties in the manual construction of concept maps.

(ii) **Information Analysis:** Verifies whether the observed results correspond to those expected. For this we used the *Statistical Analysis* to compare the propositions extracted from the text with those extracted by humans and to compare different approaches of concepts ranking; *Empirical Analysis* on the quality of the map; summarization and investigation treatment on the manual construction of the concept map from a text.

(iii) **Conclusion:** Based on the information and analysis performed during the research, we conclude that the construction of a concept map from texts is a difficult task even for human experts, and although the study did not fully resolve this issue, it presents contributions and promising results for the area.

1.8 Research Contributions

The main contributions that this research brings to the education and research community are as follows:

(i) A Categorization of Technological Approaches for Concept Maps Mining from Text;

(ii) A Model for Concept Map Mining based on four dimensions (Data Source Description, Domain Definition, Elements Identification, Map Visualization);

(iii) The ExtroutNLP[1] API containing libraries for information extraction in the Portuguese and English language, besides of services publicly available for consultation and expansion;

(iv) The HAF Model for concept ranking;

(v) The VertexSort Model for classify vertices type of a graph. The propositions extracted from the text are converted into graph and VertexSort model is used to summarize these propositions.

(vi) A Parser model for the Portuguese language in the Stanford NLP format, version 3.7.0, available in web site[2];

(vii) The CMBuilder[3] web tool for the automatic construction of concept maps from texts in Portuguese and English language;

(viii) Thesaurus, a multi-domain knowledge base, consisting of concepts and relations extracted from concept maps automatically constructed by the CMBuilder tool.

1.9 Scientific Production

(i) Aguiar, C. Z., Cury, D., & Gava, T. (2015). **Um Estudo sobre Abordagens Tecnológicas para a Geração de Mapas Conceituais**. In: XXI Congreso Internacional de Informática Educativa – TISE. Anais Nuevas Ideas en Informática Educativa, Santiago: Chile, v.11, pages 136-146, ISBN 978-956-19-0929-8.

(ii) Aguiar, C. Z., Cury, D., & Gava, T. (2015). **Uma Abordagem Tecnológica para a Construção de Mapas Conceituais.** In: XXI Congreso Internacional de Informática Educativa – TISE. Anais Nuevas Ideas en Informática Educativa, Santiago: Chile, v.11, pages 555-560, ISBN 978-956-19-0929-8.

(iii) Aguiar, C. Z., & Cury, D. (2016). **A Categorization of Technological Approaches to Concept Maps Construction.** In XI Latin American

[1] extroutnlp.lied.inf.ufes.br

[2] extroutnlp.lied.inf.ufes.br/resources

[3] cmpaas.inf.ufes.br/cmbuilder

Conference on Learning Objects and Technology (LACLO). San Carlos: Costa Rica, pages 1-9, IEEE, DOI 10.1109/LACLO.2016.7751743.

(iv) Aguiar, C. Z., Cury, D., & Zouaq, A. (2016). **Automatic Construction of Concept Maps from Texts.** Proceedings of the 7th International Conference on Concept Mapping – CMC. Innovating with Concept Mapping, Tallinn: Estonia, v.2, pages 20-30, ISBN 978-9949-29-269-1.

1.10 Organization of this Dissertation

The research developed in the course of this dissertation is divided into nine chapters. The chapters that follow this Introduction are:

Chapter 2: Explores the context of concept maps and their construction process. Besides that, we propose a model for Concept Map Mining based on four dimensions.

Chapter 3: Explores the Text Mining context including tasks related to pre-processing, extraction of patterns and analyze results steps. Presents a theoretical basis on the concepts applied in this research regarding to Natural Language Processing, Information Retrieval and Extraction, areas aimed at structuring of knowledge from unstructured text.

Chapter 4: Proposes a categorization of technological approaches for the construction of concept maps and conducts a literature review on the approaches included in this context. From the proposed categorization, we analyze and identify the related works.

Chapter 5: Describes a conceptual model for the automatic construction of concept maps as summarization of texts, i.e., the foundation of this research. The model is oriented to services and consists of four servers.

Chapter 6: Describes a technological architecture from the conceptual model defined in Chapter 5. Presents all the technological components used in the architecture, as well as its operation and integration.

Chapter 7: ExtroutNLP API is presented, a Java API for information extraction. Describes the libraries applied in the technological architecture proposed, as well as some experiments.

Chapter 8: CMBuilder is presented, a web tool for the automatic construction of concept maps as summarization of texts. Presents the tool interface, describes the process and discusses the experiments.

Chapter 9: Presents the final considerations and discusses future work.

Appendix A: Presents the questionnaire created for analysis of the manual construction of the concept maps.

Appendix B: Presents the *Quiz A* used for collecting data on the influence that the concept map automatically summarized from a text has for the understanding of the subject addressed in that text.

Appendix C: Presents the *Quiz B* used for collecting data on the influence that the concept map automatically summarized from a text has for the understanding of the subject addressed in that text.

Chapter 2
Concept Maps and their Construction Process

In this chapter we explore the context of concept maps and their construction process, which are the key issues for understanding this research. This chapter is organized as follows: Section 2.1 provides a brief introduction to concept maps; Section 2.2 discusses the concept map under the bias of representation of knowledge and information; Section 2.3 discusses and proposes a construction process of concept maps; and Section 2.4 presents some preliminary considerations of this chapter.

2.1 Concept Maps

Concept maps were proposed by Novak (NOVAK & CAÑAS, 2010) as a tool for representing and organizing knowledge, since the cognitive structure of an individual can be interpreted as a collection of concepts related with each other, in order to form significant propositions. A concept is defined as a regularity perceived in events or objects, or records of events or objects, designated by a label. A proposition is defined as a meaningful statement about an event or object. Therefore, the propositions are formed from the triple (*concept, relation, concept*) in order to constitute a semantic unit.

We are interested in the concept maps, essentially, of *scientific style*, where every concept label consists of one or more words containing a noun and every relation label consists of one or more words containing a verb. On a map, the concepts are represented by the ellipses or rectangles, and the relations are represented by a labeled directional arrow. This is the basic structure of a concept map, usually organized hierarchically, in an arborescent way. According to Ausubel's Meaningful Learning Theory (TAVARES, 2007), a mental structure of knowledge creates a meaning more efficiently when it initially considers the learning of more general and inclusive issues, rather than working with more specific issues.

Following this theory, the knowledge is assimilated by subsumers, where more general and already stable concepts contained in the cognitive structure of an individual lend themselves to anchor new and more specific concepts. For the anchoring of new concepts to be meaningful, the cognitive structure of the individual should have the necessary pre-existing concepts. Therefore, as stated by (AUSUBEL, et al., 1968), "the most important single factor influencing learning is what the learner already knows".

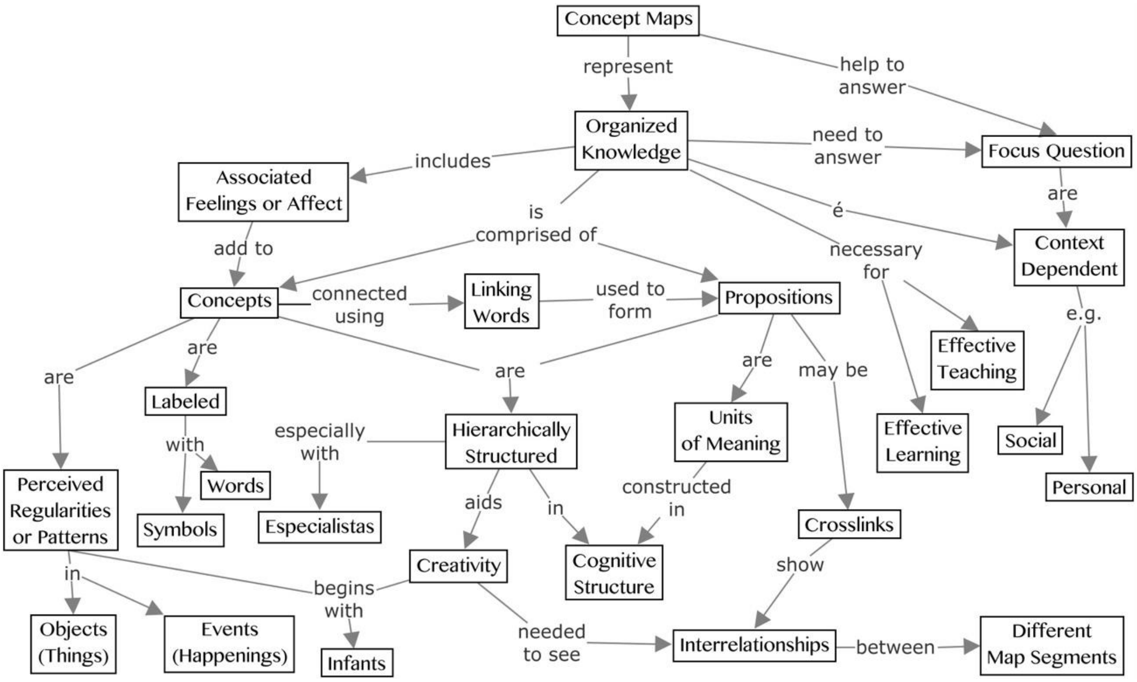


Figure 2.1 Example of Concept Map (NOVAK & CAÑAS, 2010)

Figure 2.1 shows the basic constituent elements of a concept map. Looking at the figure we note that the hierarchical organization of concepts is established by the position of elements on the map. Usually the most generic concepts appear at the top of the map, while the most specific appear at the bottom. Furthermore, the arrows may indicate the sequence and the direction of how the knowledge is built.

In addition to these characteristics, the concept map is constructed from a focal question, which organizes the relevant knowledge to answer a question in order to provide a context for the map. The map is built on a single focal question, although this issue may cover different domains or segments. Thus, cross-links are responsible for establishing explicit relationships between concepts of different or distant domains.

We can consider numerous contexts in which 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.

With regard to its **construction**, a concept map facilitates in transforming tacit knowledge into explicit knowledge, since it does not require strict formats for its representation. **Understanding** concept maps, provides one with a simple and objective way to remember pieces of information, identify relevant concepts of a domain, or view knowledge from different angles. **Sharing** maps can disseminate knowledge representation on an domain or among a group of individuals. In this case, it can be regarded as an intermediate representation of a lightweight ontology.

In this sense, concept maps have been considered a successful tool to elicit, assimilate and share knowledge in a particular domain, be it in educational or other contexts.

2.2 Representing Information using Concept Maps

Dispersion of knowledge is the main factor that creates values for society. In this context, the knowledge is created through the interaction between tacit and explicit knowledge (NONAKA & TAKEUCHI, 1997). Tacit knowledge is subjective and internalized in people's minds while explicit knowledge is transmitted by means of a formal and systematic language.

The information is a knowledge recorded in written form or oral or audiovisual, which involves an element of meaning (LE COADIC, 1996). Therefore, information must be informative, orderly or somehow structured, because otherwise it remains unusable and amorphous (MCGARRY & DE LEMOS, 1999). In this regard, explicit information promotes assimilation and interpretation, thus generating tacit knowledge.

One of the most used means for communicating information is the language, whether spoken or written. To represent 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 to the knowledge itself, the representation would require a sequential organization, adoption of a style, compliance with grammar rules, concern with format and others (GAVA, et al., 2003).

In the following, we exemplify the difference of representing information as a written text, and as a concept map (Figure 2.2). In the text, the information designating a meaning and acting as subject, object or complement of sentence 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 point out that the concept map does not represent all the information of the text, but only 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".*

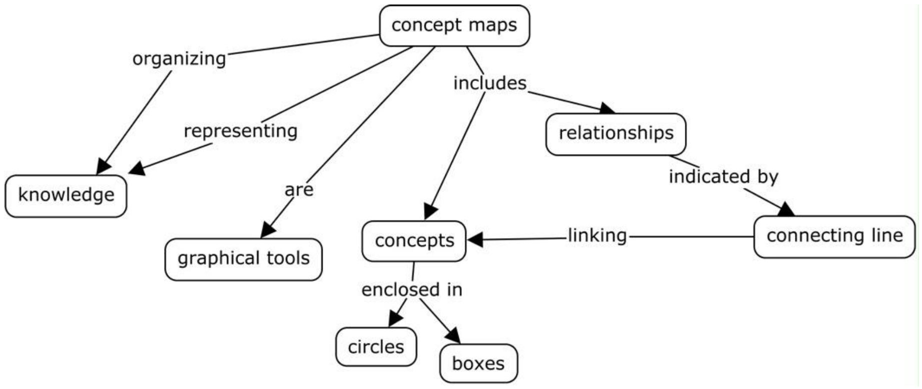


Figure 2.2 Representation of information as a written text extracted from (NOVAK & CAÑAS, 2010) and concept map constructed from it.

A text can be represented by a concept map in order to provide graphic and holistic information. In other words, this dynamic and flexible graphical representation can be considered easier to be constructed, assimilated and understood than a written text. Expert representations such as concept maps help the reader to understand text as well as to assimilate its information from a prior knowledge (PIRNAY-DUMMER & IFENTHALER, 2011).

However, we cannot consider concept map as a complete representation of the relevant logical propositions and cognitive nature expressed in a text. Yet, we can say that maps can be an integrated and meaningful representation of that cognitive nature. In addition to providing the reader with a new way to view the information, the map contributes to the discovery of new viewing angles. Therefore, we consider that the map works as a tool for the knowledge engineering. It offers a new perception on a domain, which influences the modification of pre-existing knowledge and the construction of new knowledge.

2.3 Construction Process of Concept Maps

The standard procedure for building a concept map involves (CAÑAS, et al., 2003) *(i)* defining a topic or focal question, *(ii)* identifying and listing of the most important or "general" concepts related to the topic, *(iii)* ordering the relevant concepts from top to bottom in the map and *(iv)* adding and labeling the linking phrases. Therefore, the manual construction of a concept map requires a significant amount of time and committed effort in identifying and structuring knowledge, especially when the construction of the map is performed from scratch, i.e., when its constituent elements are not predefined and they need to be fully identified.

In order to assist the construction process of concept maps, some studies have focused efforts to propose processes for their automatic construction from documents. This process is referred to as *Concept Map Mining* (CMM) (VILLALÓN & CALVO, 2011).

The generic CMM process proposed by (VILLALÓN & CALVO, 2011) can be formalized by defining a document *D* as a set $D = \{C_d, R_d\}$, where C_d is the set of all concepts, and R_d is the set of all relations extracted from the document.

This extraction process may be synthesized in the following steps: *(i)* Concept Identification, which extracts all possible concepts $C_d$from the document *D*; *(ii)* Relationship Identification, which extracts the relations R_d between two possible concepts C_d from the document *D*; *(iii)* Summarization, which reduces the map to the relevant elements for the domain, represented by $CM = \{C, R, T\}$, where the map *CM* is defined by the set of concepts *C*, relations *R* and their topological organization *T*, as shown in Figure 2.3.

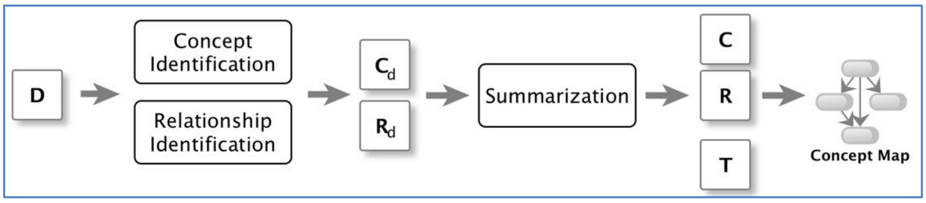


Figure 2.3 The Process of the Concept Map Mining (VILLALÓN & CALVO, 2011)

Using a different perspective, we propose a process for building maps covering four dimensions: *(i)* Data Source Description, which defines the type of data source that will be used for the construction of the map; *(ii)* Domain Definition, which identifies the domain of the data source; *(iii)* Elements Identification, which can be regarded as the core of the process making use of the earlier steps to extract concepts and relationships; and, *(iv)* Map Visualization, which specifies the graphic positioning of propositions in the concept map. After all, such dimensions should be understood as the steps for the automatic extraction of concept maps from texts, showed in Figure 2.4.

The proposed process starts with the **Data Source Description**, in order to characterize a document *D*.

In the **Elements Identification** step, a document *D* of size *n* can be defined as

$D = \{d_1 \ldots d_n\}$ *where* $d_i, i=1\ldots, n$ is a term in *D*.

A set of concepts can be defined as

$C = \{c_1 \ldots c_n\}$ *where* $C \subseteq D$

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

A set of relationships can be defined as

$R = \{r_1 \ldots r_n\}$ *where* $R \subseteq D$

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

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

A proposition can be defined as $P_{ijk}=\{c_i,r_j,c_k\}$ where $c_i \in C$ and $c_k \in C$ and $r_j \in R$.

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

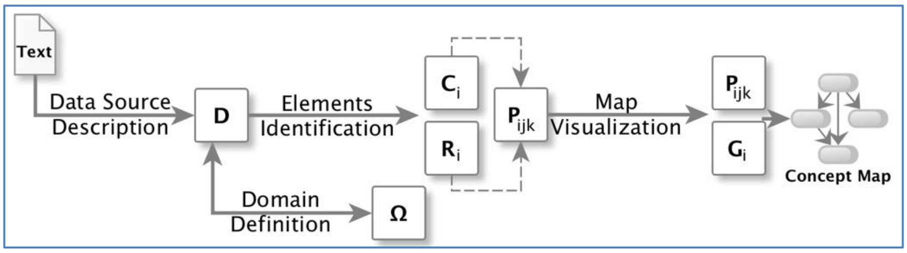


Figure 2.4 Process of the Concept Map Mining proposed (by Author)

The **description of the data source** impacts the whole process of building the concept map. In this step, we define some characteristics, especially with respect to the size and quantity of information available in the data source. For the size, we can characterize unstructured data sources as *(i)* small: small content such as abstracts; *(ii)* regular: few data pages such as academic articles, reports, newspaper, articles etc.; *(iii)* long: extensive data containing a lot of information such as theses and dissertations. According to the quantity, the data source may be represented in two groups: *(i)* approaches that use a set of documents to represent the knowledge of a domain and *(ii)* approaches that use a single document that represents the knowledge specific to one author.

We believe that one of the challenges for the automatic construction of concept maps from texts is the **definition of the domain**, i.e. of the text domain or the concepts belonging to the domain. In this context, we note the use of semi-automatic techniques where the author identifies the domain of the data source by choosing a suitable ontology (GRAUDINA & GRUNDSPENKIS, 2008), or using multiple maps (VALERIO, et al., 2008), or using a list of concepts (CLARIANA & KOUL, 2004) or by means of a set of documents (LAU, et al., 2007).

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

The **map visualization** step shows the topological structure of propositions identified in the elements identification step by means of a graphical interface. In this case, we observed that many approaches use outsourced tools for these purposes (VILLALÓN & CALVO, 2011) (CLARIANA & KOUL, 2004). However, some approaches develop their own display interface including features that facilitate learning, such as a list of occurrences of the concept within the context (ZOUAQ, et al., 2007), a partial map view from the perspective of a concept (LAU, et al., 2007), or the display of the path of a specific concept until the focus question (KUMAZAWA, et al., 2009).

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

2.4 Some Considerations on the Chapter

To introduce some key issues to understanding this research, this chapter briefly broached issues on concept maps, their underlying theories, representation and construction.

Regarding the construction of concept maps, the CMM process has been covered and discussed in great detail, according to the version proposed by (VILLALÓN & CALVO, 2011). This version has been extended as a new proposal for the CMM process based on four dimensions of interest.

The next chapter introduces approaches and techniques based on text mining, necessary for the technical understanding of this work.

Chapter 3
Text Mining and Information Extraction

In this chapter we explore the context of Text Mining (TM) and Information Extraction (IE), two areas which aimed at extracting knowledge from unstructured text. Although the Information Extraction area is included in the Text Mining area, we emphasize its importance for the development of our research.

This chapter is organized as follows: Section 3.1 introduces concepts of the Text Mining area, the steps of the text mining process are included in the following subsections: Section 3.1.1 Pre-Processing Step, Section 3.1.2 Patterns Extraction Step and Section 3.1.3 Evaluation Step; Section 3.2 introduces concepts of the Information Extraction area and their main techniques; and Section 2.4 presents some preliminary considerations of this chapter.

3.1 Text Mining

Text mining (TM) is characterized by providing an interpretable information from unstructured data, i.e., it refers to the process of information extraction or knowledge discovery from textual documents. Text mining can be considered as a subfield of data mining. The first one identifies implicit and useful information on the unstructured data, and the second tries to find interesting patterns from large structured data like databases.

Text mining is an interdisciplinary field that incorporates areas such as information retrieval, information extraction, data mining and natural language processing (SUMATHY & CHIDAMBARAM, 2013), as shown in Figure 3.1 and explained in the following.

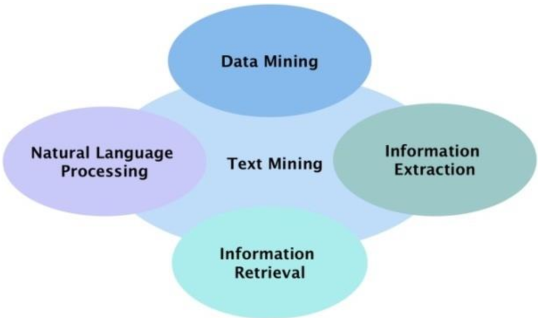


Figure 3.1 Text mining areas

Natural Language Processing (NLP): By "natural language" we mean a language that is used for everyday communication by humans, such as English or Portuguese. The NLP, also called Computational Linguistics, is an attempt to achieve a better understanding of natural language by use of computers (KODRATOFF, 1999), either by interpretation or generation of natural language. At one extreme, it can be as simple as counting word frequencies to compare different writing styles. At the other extreme, NLP involves "understanding" of complete human utterances, at least to the extent of being able to give useful responses to them (BIRD, et al., 2009).

Information Retrieval (IR): Information Retrieval deals with the representation, storage, organization, and access to information items such as documents (BAEZA-YATES & RIBEIRO-NETO, 2013), making large volumes of text accessible to people with information needs (SALTON & MCGILL, 1983). IR is used to find a document of an unstructured nature within large collections of documents, which are processed to condense or extract the particular information sought by the user.

Information Extraction (IE): Information Extraction finds and connects relevant information and, at the same time, ignores different or irrelevant information in an unstructured document (COWIE & LEHNERT, 1996), i.e. it extracts specific information in a structured format.

Data Mining (DM): Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. Data is characterized as recorded facts and Information is characterized as the set of patterns, or expectations, that underlie the data (WITTEN & FRANK, 2005). In practice, the two primary goals of data mining tend to be prediction and description. *Prediction* uses some variables in the data to predict unknown or future values of other variables of interest, while *Description*, finds patterns describing the data that can be interpreted by humans (KANTARDZIC, 2011). The process must be automatic or semi-automatic and the discovered patterns must be meaningful.

Text mining involves a set of computational methods used to navigate, organize, find and discover information in textual data that normally could not be retrieved using traditional methods. The goal of text mining is to discover relevant information in the text by transforming the text into data that can be used for further analysis.

The text mining process proposed by (REZENDE, 2003) and shown in Figure 3.2 will be adopted by this research.

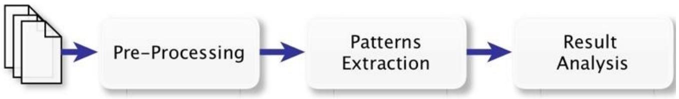


Figure 3.2 Text mining process

The first step of the text mining process is defining which documents, a.k.a. corpus, will constitute the input data source. Such documents must be relevant to the domain and may refer to single or to set of documents. These steps are described in the following sections.

3.1.1 Pre-Processing Stage

Pre-Processing stage is an important task for the text mining process and it is crucial in determining the quality of the next stages, since it selects only the significant keywords. The overview of pre-processing tasks is depicted in the Figure 3.3 and explained as follows:

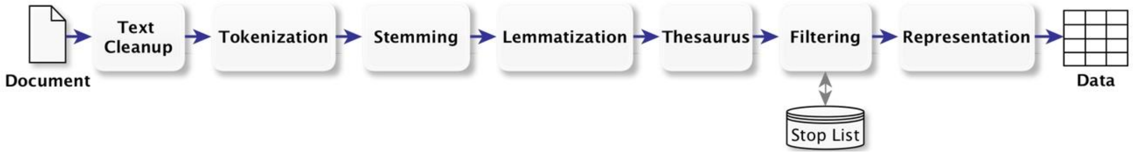


Figure 3.3 Pre-Processing tasks

3.1.1.1 Text Cleanup

Text Cleanup is the character removal task that does not contribute to the knowledge extraction, i.e., the "noise". These characters may be invalid and/or not belong to a set of letters, numbers, special characters, punctuation, and others, for instance ( ) § # | { } @.

3.1.1.2 Tokenization

Tokenization or lexical analysis is the process of converting a sequence of characters (text) into a sequence of meaning units (words) that compose the text. The term token is used to designate these meaning units, which correspond to one or more textual units such as "27/01/2017", "UFES", "100,00" and "pre-processing". The tokenization process is performed by delimiter tokens such as characters or formatting controls. The spaces and punctuation are generally adopted as delimiter tokens for Western languages (FELDMAN & SANGER, 2007).

Although this task may seem easy for humans, it is complex for machines. According to (GASPERIN & LIMA, 2000), some challenges are related to: *(i)* punctuation, as it may indicate the phrase end, an abbreviation, or a formatting; *(ii)* apostrophe, as it may indicate a contraction or possessive case; and *(iii)* hyphen, as it may indicate a compound word, syllable separation, or word qualification; among others.

To clarify, the following sentence "*This phrase is an example of tokenization.*" can be represented with the following tokens sequence: *[this] [phrase] [is] [an] [example] [of] [tokenization] [.]*.

3.1.1.3 Lemmatization

This method is used to find the lemma of the word, the base form, disregarding grammatical changes such as tense and plurality (BIBER, et al., 1998). Lemmatization is representing the word in its canonical form. The canonical form for verbs is the infinitive, and for adjectives and nouns the masculine singular (ARAMPATZIS, et al., 1999).

For instance, the set of terms *connect*, *connected* and *connecting* can be represented by the common lemma *connect* and the set of terms *connection* and *connections* can be represented by the common lemma *connection*.

3.1.1.4 Thesaurus

Besides the lemmatization, thesaurus can also be considered a good strategy for the reduction of the dimensionality, since it organizes the semantic value of terms using the mapping of synonymous, hierarchies and relationships (EBECKEN, et al., 2003). A thesaurus is a controlled vocabulary, formally organized and where the a priori relationship between concepts is explicit (AITCHISON, et al., 2000).

A thesaurus is created for different contexts, among others to represent lexical information between synonymous words, such as WordNet (FELLBAUM, 1998), or to represent the relationship between the items of information within a domain knowledge, such as the clinical (SIOUTOS, et al., 2007), ecological (MAGGIORE & ANZALDI, 1998), and engineering-to-biology (STROBLE, et al., 2009) domains.

According to (FOSKETT, 1997), the main purposes of a thesaurus are basically: *(i)* to provide a standard vocabulary for indexing and searching; *(ii)* to assist users with locating terms for proper query formulation; and *(iii)* to provide classified hierarchies that allow the broadening and narrowing of the current query request according to the needs of the user.

3.1.1.5 Filtering

The filtering removes irrelevant words, meaning an attempt to remove all information that does not constitute knowledge in the text. The standard filtering method is the removal of stop-words, based on a set of irrelevant words called stop-list. The idea of this filtering is to remove words that bear little or no content information, such as "*this*", "*in*", "*a*", "*an*", "*with*", "*of*", among others. Typically, 40 to 50% of the total words in a text are removed with a stop-list (SALTON & MCGILL, 1983). Furthermore, terms that occur with high frequency or occur rarely are probably not of great relevance and can be removed (FRAKES & BAEZA-YATES, 1992).

For instance, the following sentence "*This phrase is an example of tokenization.*" can be represented after stop-words filtering with the following tokens sequence: *[phrase] [is] [example] [tokenization] [.]*.

3.1.1.6 Document Representation

An important activity that must be performed during pre-processing stage is the choice of how to represent the terms of the text. A textual document is formed by a collection of words and their occurrences. This allows for the transformation of the information into a structured format, usually in numerical representation. In this representation, the data represents an economic and meaningful way to be analyzed and processed, such as Vector Space Model (VSM).

The VSM represents documents as numerical vectors, i.e., it forms a matrix of high dimension for the respective document. The simplest representation of texts introduced within vector model is called "bag of words" (SALTON & MCGILL, 1983), in this case the occurrence order of each word in the document is not considered. The success or failure of the vector space model is based on the term weighting (POLETTINI, 2004).

For a document collection $D = \{d_1, \ldots, d_n\}$ and the respective terms $K = \{k_1, \ldots, k_n\}$, a weight $W_{i,j}$ is assigned to the term-document pair (k_i, d_j). The weight $W_{i,j}$ can be calculated using different types of weighting, as shown below.

Term Frequency (TF): is based on the assumption that the weight of a term occurring in a document is directly proportional to its frequency (LUHN, 1957). The weight $W_{i,j}$ for the frequency $tf_{i,j}$ of a term k_i occurring in a document d_j can be defined by Formula 3.1.

$$W_{k,d} = \begin{cases} 1 + \ log_{10} tf_{i,j} & if\ log_{i,j} > 0 \\ 0 & \quad otherwise \end{cases} \qquad (3.1)$$

Inverse Document Frequency (IDF): is based on the assumption that the specificity of a term can be measured by an inverse function of documents number in which it occurs (SPARCK JONES, 1972). Since DF is the frequency of documents d_i that the term k_i occurs, IDF is its inverse frequency. The weight IDF_i for the documents frequency df_i of a term k_i occurring in a documents collection N can be defined by Formula 3.2.

$$IDF_i = log \frac{N}{df_i} \qquad (3.2)$$

Term Frequency and Inverse Document Frequency (TF-IDF): is one of the most popular weighting schemes and combines the TF and IDF factors (SALTON & YANG, 1973). The weight $W_{i,j}$ associated to the term (k_i, d_j) can be defined by the Formula 3.3.

$$W_{i,j} = \begin{cases} (1 + log\, f_{i,j}) \times log \frac{N}{df_i} & if\ f_{i,j} > 0 \\ 0 & otherwise \end{cases} \qquad (3.3)$$

3.1.2 Patterns Extraction Stage

After the pre-Processing stage, algorithms and techniques of data mining are applied to extract knowledge. At this point the Text Mining (TM) process merges with the traditional Data Mining (DT) process. The first process works on text and the second one on structured data.

The choice of the techniques for finding and describing structural patterns depends on how the extracted knowledge will be interpreted, as well as on the computational time required and the purpose of the approach. Some main tasks of the data mining process will be discussed in the following; the information extraction task will be addressed in Section 3.2.

3.1.2.1 Clustering

Clustering is a descriptive task in which one seeks to identify a finite set of *clusters* to describe the data (KANTARDZIC, 2011) based on associating among features within the data and on the contexts they have in common (STRZALKOWSKI, 1999). Clustering is the most common unsupervised learning task, as it does not assume the existence of a teacher for estimating the proposed model.

Given a document collection $D=\{d_1,\ldots,d_n\}$, a textual clustering method automatically separates these documents into clusters $K=\{k_1,\ldots,k_n\}$ according to some predefined criterion (BAEZA-YATES & RIBEIRO-NETO, 2013). This criterion is usually adopted by the degree of similarity or dissimilarity between the documents. The minimum Euclidean distance is equivalent to the maximum Cosine similarity.

Euclidean Distance: is the most well-known distance measure, presented in Formula 3.4 and titled as d_{Euc}. In the formula, *P* and *Q* are vectors of the terms from the two documents and the value d_{Euc} closest to 0 indicates similar documents.

$$d_{Euc} = \sqrt{\sum_{i=1}^{d} |P_i - Q_i|^2} \qquad (3.4)$$

Cosine Similarity: is defined from the Euclidean n-dimensional space model and presented in Formula 3.5, titled as s_{Cos}. In the formula, *P* and *Q* are vectors of terms from the two documents and the value s_{Cos} closest to 1 indicates similar documents.

$$s_{Cos} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sqrt{\sum_{i=1}^{d} P_i^2}\sqrt{\sum_{i=1}^{d} Q_i^2}} \qquad (3.5)$$

There are many ways to create clusters, however the most ways are variations on a few basic algorithms (WILLETT, et al., 1998). One of the widely known, simple and effective algorithms is the K-Means (MACQUEEN, 1967).

K-Means: this method defines in advance how many clusters are being sought, the parameter *k*. Then *k* points are chosen as cluster centroid. The centroid is a subset or center point of a cluster. Then, each document of the collection is assigned to the closest centroid according to the Euclidean distance. The following shows the main steps of the K-Means algorithm:

(i) Randomly select *k* centroids.

(ii) Calculate the distance between each data point and centroids.

(iii) Attribute the closest cluster to each data point.

(iv) When all data-points have been assigned, recalculate the new centroids.

(v) Recalculate the distance between each data point and new obtained centroids.

(vi) If no data point was reassigned then stop, otherwise repeat from step *(iii).*

3.1.2.2 Classification

Text classification provides a means to organize information allowing for a better understanding and interpretation of the data (BAEZA-YATES & RIBEIRO-NETO, 2013). The set of documents whose contents can be described by a label is called *class*. The classes are arranged in a hierarchy or network reflecting the concepts that define the domain of the corresponding document collection (STRZALKOWSKI, 1999). A label can be a topic, such as finance and sports, or a genre, such as news and movie, or an opinion, or domain-specific.

Given a document collection *D* and a set of classes *C* with their respective labels, a textual classifier assigns a class for each pair $[d_i, c_j]$ according to a metric, such as probability and similarity. One of the oldest and simplest classification methods is the K-Nearest Neighbor (COVER & HART, 1967),

K-Nearest Neighbor (KNN): is a lazy learning classifier that builds the classification model only when a new document is submitted (BAEZA-YATES & RIBEIRO-NETO, 2013). The algorithm is based on the distance function for pairs of observations. The classification decision is based on the classes of *k* closest neighbors of the document. The following shows the main steps of the KNN algorithm:

(i) The distance between the document d and each training document is calculated using some similarity measure such as the Cosine measure.

(ii) The k closest training documents are selected, i.e. documents more similar to the document d.

(iii) The document d is classified in a category according to some grouping criterion defined in the training documents.

3.1.2.3 Summarization

Summarization is a brief and accurate representation of an input text of the type that the output covers the most important concepts of the data source in a condensed manner (THAKKAR, et al., 2010). According to (HUTCHINS, 1987), scientific summaries can be classified into three types: *(i) indicative*, containing only the essential topics of a text; *(ii) informative*, containing all the main aspects of the text and considered as a replacement for the text; and *(iii) evaluative*, presenting a comparative analysis between the content of the text source and other related works.

This research is interested in scientific summaries of informative type. Therefore, the automatic text summarization is a task that creates a compact representation of a document or documents collection for understanding and covering its main purpose.

There are two main approaches to the summarization task, which are *extraction* and *abstraction* (HAHN & MANI, 2000). *Abstraction* is a summary produced by reformulating sentences (TORRES-MORENO, 2014), i.e., it interprets the information contained in the original source and generates a text that expresses the same information in a more concise way. *Extraction* is a summary produced by extracting sentences from the text source (TORRES-MORENO, 2014), i.e., it selects pieces of text (words, phrases, sentences, paragraphs) from the original source organizing them in a way to produce a coherent summary. Although a high-quality abstraction-based summarizer will potentially be more useful, the research in automatic summarization is mainly focused on extraction-based methods because they employ a more straightforward approach for constructing summaries (SIZOV, 2010).

Thus, an essential part of the extraction-based approach is the identification of sentences containing important information (SIZOV, 2010), in order to detect the content that should be kept in the summary. It can be done using graph-based representations by means of ranking algorithms.

The graph represents the text, where the vertices are text units (word, collocations, sentence etc.) and edges interconnects vertices with meaningful relations. Ranking is essentially a way of deciding the importance of a vertex within a graph based on information drawn from the graph structure (THAKKAR, et al., 2010). From it is possible to find more representative keywords or phrases to build the summarization.

In this context, two graph-based ranking algorithms are given importance in the literature, Hyperlink-Induced Topic Search (KLEINBERG, 1999) and PageRank (PAGE, et al., 1999). The algorithms were developed in the link structure context of the web in order to discover and rank relevant pages to a particular topic. However, the same idea presented to internet pages can be used for the representation of text.

The algorithms assign an arbitrary value to each vertex (page) in the graph which then iterates until convergence below of a given threshold. Finally, a score is associated with each vertex, which represents the "importance" of that vertex within the graph.

Hyperlink-Induced Topic Search (HITS): determines two values for a page: its *authority*, which estimates the value of the number of incoming links, and its *hub* value, which estimates the value of its links to other pages. The authority value is defined as shown in Formula 3.6, where $V_j = \{v_1 \dots v_n\}$ are pages linking to page V_i. The hub value is defined as shown in Formula 3.7, where $V_j = \{v_1 \dots v_n\}$ is the number of outgoing links from a page V_j (out-degree).

$$HITS_{Aut}(V_i) = \sum_{V_j \in In(V_i)} HITS_{Hub}(V_j) \qquad (3.6)$$

$$HITS_{Hub}(V_i) = \sum_{V_j \in Out(V_i)} HITS_{Aut}(V_j) \qquad (3.7)$$

Authority and hub values are defined in terms of one another in a mutual recursion in *k* iterations. The hub score and authority score for a node is calculated with the following algorithm (THAKKAR, et al., 2010):

(i) Start with each node having a hub score and authority score of 1;

(ii) Run the Authority Update Rule;

(iii) Run the Hub Update Rule;

(iv) Normalize the values by dividing each Hub score by the sum of the squares of all Hub scores, and dividing each Authority score by the sum of the squares of all Authority scores;

(v) Repeat from the second step as k iterations.

PageRank: this algorithm integrates the impact of both links, incoming and outgoing, into one single model. The value is defined as shown in Formula 3.8, where $V_j=\{v_1 \dots v_n\}$ are pages linking to page V_i, $Out(V_j)$ is the number of outgoing links from a page V_j (out-degree) and *d* is a damping factor between 0 and 1, usually set to 0.85.

$$PR(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|} \qquad (3.8)$$

3.1.3 Results Analysis Stage

This stage is responsible for the evaluation and interpretation of the results. *Subjectively*, the results can be evaluated by the end user, domain expert or data analyst, in order to validate the knowledge (EBECKEN, et al., 2003). *Objectively*, the results can be estimated using an approximate metric, for instance, which statistically compares the results produced by other approaches.

Thus, efficient information retrieval depends on two main factors (SALTON & BUCKLEY, 1988): *(i)* relevant items must be retrieved; and *(ii)* non-relevant items must be rejected. For this, the statistical metrics of precision, recall and f-measure are commonly used.

Precision: is defined as the amount of correctly extracted information of the total existing information in the text, i.e., proportion of retrieved items that are relevant (BAEZA-YATES & RIBEIRO-NETO, 2013). This metric is defined as follows in Formula 3.9.

$$precision = \frac{Number\ of\ actual\ slot\ values\ correctly\ predicted}{Number\ of\ slot\ values\ predicted\ to\ be\ present} \qquad (3.9)$$

Recall: is defined as the amount of correctly extracted information of all relevant information from the text, i.e., it defines how complete or comprehensive the extraction of relevant information is (HOBBS, et al., 1997). This metric is defined as follows in Formula 3.10.

$$recall = \frac{Number\ of\ actual\ slot\ values\ correctly\ predicted}{Number\ of\ actual\ slot\ values} \qquad (3.10)$$

In practice, precision and recall tend to vary inversely, since it is very difficult to recover everything that is relevant and remove everything that is not relevant. Figure 3.4 represents the precision and recall for a given information extraction *I* (BAEZA-YATES & RIBEIRO-NETO, 2013).

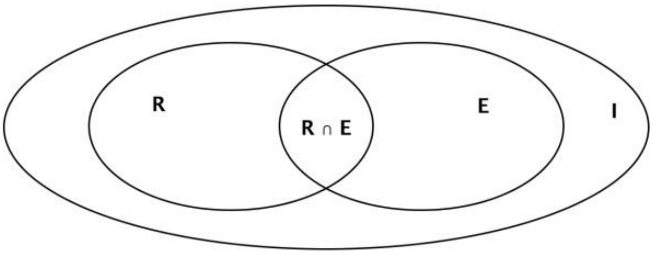


Figure 3.4 Precision and recall for an information extraction I (BAEZA-YATES & RIBEIRO-NETO, 2013) (adapted)

The total information from a collection is represented as I, the subset of relevant information is represented as R, the subset of information extracted is represented as E, and the intersection between the sets is represented as $R \cap E$. Thus, precision is defined as $R \cap E/E$ and recall is defined as $R \cap E/R$.

F-measure: is the metric that combines recall and precision measurements within a single value (HOBBS, et al., 1997), defined as follows in Formula 3.11.

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \qquad (3.11)$$

3.2 Information Extraction

Much of the work in text mining makes uses of statistical-based methods, treating documents as an unordered bag of words or vector space model as it is typical in information retrieval (MOONEY & BUNESCU, 2005). This representation type has been shown to be effective in a number of information retrieval tasks. However, in linguistic methods the knowledge that might be mined from text includes identifying entities, properties and relationships between elements of the text.

Although full natural language understanding is still far from the capabilities of current technology, existing methods in Information Extraction (IE) are able to recognize several types of elements in the text and identify some relationships that are asserted between them (MOONEY & BUNESCU, 2005). Therefore, IE can serve as an important technology for text mining.

The IE extracts specific information within an unstructured textual document, and then the information is structured in a tabular form. IE does not interpret the text in all its parts; instead it analyzes parts of the text that contain relevant information to specific domain (CORRÊA, 2003). Some of the main tasks related to IE process are discussed in the following.

3.2.1 Morphological Analysis

Morphological Analysis is focused on the individual terms. For each word in a sentence the analysis identifies its grammatical class or part of speech (noun, verb, preposition etc.) and its flexion (gender, number and grade). A morphological tagger, pos-tagger, assigns specific tag to words according to their grammatical class.

The set of tags that have been used for the English language is the Penn Treebank tag (SANTORINI, 1990). For the Portuguese language, there is no consensus on the set of tags used, however we can highlight Forest Treebank (AFONSO, et al., 2002) and Cintil Treebank (BRANCO, et al., 2010).

The most widespread corpora with annotated POS tags in Portuguese are Mac-Morpho (ALUISIO, et al., 2003), with around one million words based in proprietary tags; Bosque, with around 185 thousand based in Forest Treebank; and CINTIL-Corpus International Portuguese, with around one million words based in Cintil Treebank. Table 3.1 shows the morphological tags that will be used in the course of this research.

Grammatical Class	***Penn***	***MacMorpho***	***Cintil***
Conjunction	CC	KC	CJ
Numeral, cardinal	CD	NUM	CARD
			DGT
Determiner	DT	ART	DA
			DEM
			IA
Preposition	IN	PREP	P
	TO	PREP+ART	
Interjection		IN	ITJ
Adjective	JJ	ADJ	ADJ
	JJR		
	JJS		
Noun	NN	N	CN
	NNP	NPROP	
	NNPS		
	NNS		
Pronoun	PRP	PROPESS	PRS
		PROADJ	POSS
		PROSUB	
Adverb	RB	ADV	ADV
Verb	VB	V	INF
	VBD	VAUX	INFAUX
	VBG	PCP	VAUX
	VBN		V
	VBP		GER
	VBZ		GERAUX
			PPA
			PPT
Punctuation	.	PT	PNT

Table 3.1 The morphological tags used in the course of this research

One of the difficulties of the morphological analysis task is the existence of many words with different possible classifications, causing ambiguity. Consider the phrases "*I have a lot of work to do today.*" and "*A person must work hard to achieve his or her goal.*", the word "*work*" adopts the substantial role in the first sentence and the verbal role in the second sentence.

3.2.2 Syntactic Analysis

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 (FELDMAN & SANGER, 2007).

Constituency grammars describe the syntactical structure of sentences according to sequences of syntactically grouped elements (noun phrases, verb phrases, prepositional phrases, adjective phrases, and clause). Therefore, a noun phrase of the constituency grammar may be labeled as the subject, direct object, or the complement of a sentence. In a constituency parse tree as shown in Figure 3.5, the *non-terminal node* is the type of the phrase, the *terminal node* is the word in the sentence, and the *edge* is unlabeled.

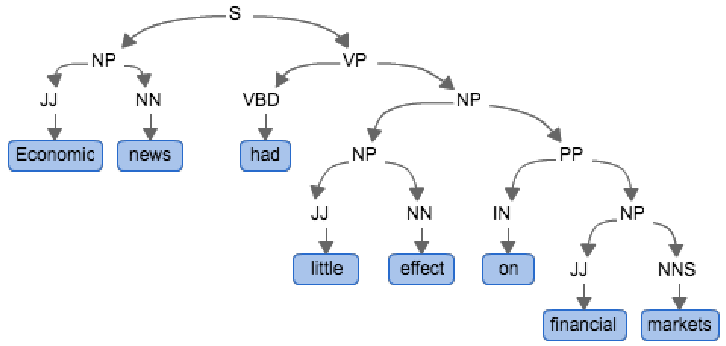


Figure 3.5 Constituent structure from Penn Treebank

Dependency grammar does not recognize the constituents as separate linguistic units. Instead, it focuses on the direct relationships between words (subject, direct object etc), i.e., it connects words according to their relationships. Thus, a subject and direct object nouns of a sentence depend on the main verb of the dependency grammar. In a dependency parse tree as shown in Figure 3.6, each *vertex* represents a word, *child node* is dependent word of the parent, and *edge* is labeled by the relationship.

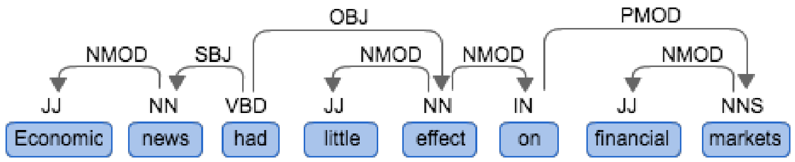


Figure 3.6 Dependency structure from the Penn Treebank

Table 3.2 shows the syntactical tags that will be used in the course of this research.

Class	*Penn*	*MacMorpho*	*Cintil*
Declarative clause	S	S	S
	SBAR		
Noun Phrase	NP	NP	NP
Verb Phrase	VP	VP	VP
Prepositional Phrase	PP	PP	PP

Table 3.2 Syntactical tags defined by the TreeBanks

3.2.3 Semantic Analysis

Semantic analysis is a process of mapping sentences in order to represent their meaning, i.e., provides common-sense knowledge about the world (CHARNIAK & MCDERMOTT, 1998).

Semantic analysis finds out the meaning of linguistic input and constructs meaning representations (DHURIA, 2015). To extract data and construct models of the world, the semantic analysis uses some approaches such as predicate logic.

The semantic analysis as a study of meaning covers the most complex tasks, including finding synonyms, word sense disambiguation, translating from one natural language into another, and populating base of knowledge, among its other functions (POROSHIN, 2014).

3.2.3.1 Semantic Similarity

The semantic similarity between terms can be calculated by several functions and diverse information processes such as Knowledge-Based Similarity. The Knowledge-Based Similarity measures calculates the degree of similarity between words using information derived from semantic networks (MIHALCEA, et al., 2006), such as WordNet (FELLBAUM, 1998). Such metrics can be based on the information content such as LIN measure (LIN, 1998).

LIN Measure: returns the information content (IC) of the least common subsumer (LCS) between two concepts. Therefore, the more information two words have in common, the more similar they are.

The LIN measure is defined as follows in Formula 3.12, where *IC(c)* is defined in Formula 3.13 and *P(c)* is the probability of encountering an instance of concept *c* in a large corpus. The result value belongs to the range of 0 to 1.

$$Sim_{lin} = \frac{2*IC(LCS)}{IC(concept_1)+IC(concept_2)} \quad (3.12)$$

$$IC(c) = -\ log\, P(c) \quad (3.13)$$

3.2.3.2 Named Entity Recognition (NER)

A named entity is a sequence of words that designates a real-world entity, such as “Brazil,” “UFES” or “Steve Jobs”. NER identifying uses rigid designators from the text belonging to predefined types, such as person, organization and location (NADEAU & SEKINE, 2007).

Solutions to named entity recognition have used Rule-based approach. This system consists of a collection of rules with hand crafted grammars or learnt from examples. Another system is Machine Learning-based approach, which is provided with a set of pre-classified (labeled) texts for each category used as the training set, and it automatically produces a classifier from them. Additionally, there are Hybrid approaches, that combine both systems.

Considering the phrase "*The Brazil Institute was created with the purpose of managing and performing public procurement across Brazil.*" the entity "*Brazil*" in the first occurrence should be recognized as an organization and in the second one as a place.

3.2.3.3 Co-reference Resolution

Co-reference resolution is the task of finding all expressions that refer to the same entity in a discourse (LEE, et al., 2013), i.e., determining which noun phrases (NPs) refer to each real-world entity mentioned in the document.

In co-reference resolution, it is common that the candidates compete to be the antecedent of an anaphor (MITKOV, 2014). Therefore, the cohesion which points back to some previous item is called anaphor (HASAN & HALLIDAY, 1976) and the entity to which it refers to or for which it stands is its antecedent.

The process of determining the antecedent of an anaphor is called anaphora resolution. When the anaphor refers to an antecedent and when both have the same referent in the real world, they are termed co-referential (MITKOV, 2014).

Considering the phrase "*The Queen is not here yet but she is expected to arrive in the next half an hour.*", the pronoun "*she*" is the anaphor, "*the Queen*" is the antecedent and both are co-referential (HUDDLESTON, 1984).

3.2.3.4 Relation Extraction (RE)

Relation Extraction (RE) is the task of detecting and characterizing the semantic relations between entities in a text (DODDINGTON, et al., 2004). Semantic relations are meaningful associations between two or more concepts, entities, or sets of entities (KHOO & NA, 2006). In the literature, RE can be applied to closed-domain and open-domain (FARUQUI & KUMAR, 2015).

The Closed-domain considers only a closed set of relationships between two arguments or entities, i.e., pre-defined and binary relationships. Various techniques, such as feature-based and kernel-based, have been proposed (NGUYEN, et al., 2015). However, these techniques, for the most part, require a large amount of training data. They are usually domain dependent, and their adaptation to a new domain requires manual labor comprising specification and implementation of new patterns of relationships or corpora annotation (EICHLER, et al., 2008). Moreover, this approach is not scalable to corpora with a large number of target relationships or where the target relationships cannot be specified in advance (ETZIONI, et al., 2011).

The phrase "*Mary works at Google headquarters in Brazil.*" informs relations such as Headquartered-in(*Google, Brazil*) and Employment(*Mary, Google*).These relations relate Organization with Location class and Person with Organization class.

On the other hand, Open-domain uses an arbitrary phrase from sentences to specify a relationship. Thus, Open RE or Open Information Extraction is a domain independent approach and does not specify the relationships in advance.

3.2.3.5 Open Information Extraction (Open IE)

Open Information Extraction is a domain-independent extraction paradigm that uses some generalized patterns to extract all the potential relationships between entities (LI, et al., 2011). Open IE aims to obtain a shallow semantic representation of natural language text as a set of triples in form of (*arg1*, *rel*, *arg2*), where *arg1* and *arg2* are noun phrases and *rel* is a textual fragment indicating an implicit semantic relationship between the arguments (WU & WELD, 2010). Each triple extracted is called proposition. Most of the Open IE techniques do not require any background knowledge or manually labeled training data; they are therefore not limited to a set of pre-specified relations or entities.

Open IE considers that all connections among concepts, entities, events, and also those expressed by means of attributes can be considered as relations (XAVIER, et al., 2013). For instance, the sentence "*Mary bought a beautiful home.*" informs relations such as *(Mary, bought, a beautiful home)* and *(Mary, bought, a home).*

Open IE makes use of hand-crafted extraction heuristics or automatically constructed training data to learn extractors of propositions (DEL CORRO & GEMULLA, 2013). For this, the approaches use shallow syntactic parsing or dependency parsing, sometimes applying identification of clause (a part of a sentence that expresses some coherent information), extraction rules and inference rules.

Concerning approaches based on syntactic parsing, we would like to single out the TextRunner system (BANKO, et al., 2007) the precursor of Open IE paradigm. TextRunner uses a small set of hand-written rules to label training data from sentences and uses a classifier to apply the extraction in online sentences. WOE (WU & WELD, 2010) uses heuristics between values of attributes in the Wikipedia and sentences to create training data, then uses a classifier to apply the extraction. Reverb (FADER, et al., 2011) is based on simple heuristics that identify verbs expressing relationships and then obtains their arguments. R2A2 (ETZIONI, et al., 2011) uses hand-labeled training data to identify the arguments of a verbal phrase by means of classifiers, the arguments are extracted along with noun phrases.

Concerning approaches based on dependency parsing, we would like to mention the DepOE (GAMALLO, et al., 2012) which uses rule-based analyzer on the dependency parsing, proposing the extraction of relations in other languages and identification of clause constituents. OLLIE (SCHMITZ, et al., 2012) uses a training data to learn extraction patterns on the dependency parsing and then applies them over the corpus. ClausIE (DEL CORRO & GEMULLA, 2013) uses a set of clauses over the dependency parse and a small set of domain-independent lexical. It uses a classification method to identify arguments of a relation, handles non-verb relations like appositions, and treats possessives.

Considering the phrase "*Concept maps include concepts, usually enclosed in circles or boxes of some type, and relationships between concepts indicated by a connecting line linking two concepts.",* the following propositions can be extracted by the ClausIE system: (*Concept maps, include, concepts usually enclosed in circles of some type*), (*Concept maps, include, concepts usually enclosed in boxes of some type*), (*concepts, be enclosed, usually in circles of some type*), (*concept, be enclosed, usually in boxes of some type*), (*concepts, be enclosed, usually*), (*relationships between concepts, indicated, by a connecting line linking two concepts*), (*relationships between concepts, indicated, linking two concepts*).

Many issues in Open IE are still far from being completely addressed, such as identifying the arguments from relations and accuracy when extracting a large number of relations from the same sentence. Regarding the failure to extract the relations between the arguments, the two most significant problems are (XAVIER, et al., 2013): incoherent extractions, that do not have a meaningful interpretation, and uninformative extractions, that do not express information.

Thus, according to (ETZIONI, et al., 2011), there are three key points that must be addressed to improve the results of Open IE techniques: *(i)* Extracting of n-ary relations, since not all relationships expressed in a text are binary; *(ii)* Learning relationships that are not expressed by verbs; and *(iii)* Extending Open IE systems to other languages than English.

3.3 Some Considerations on the Chapter

This chapter introduced key concepts for the understanding of this research with respect to Text Mining and Information Extraction. We emphasize the importance of the domain-independent extraction paradigm, since this paradigm will be the core of this research, that is, it will extract propositions from the text for the construction of the concept map.

The next chapter presents a literature review on the area and a characterization of technological approaches for the construction of concept maps is elaborated.

Chapter 4
Technological Approaches for Concept Maps Mining from Texts: Categorization and Literature Review

In this chapter we present a categorization of technological approaches for the construction of concept maps in the literature of the area between the years 1994 and 2016. The categorization is applied in order to provide a greater objective analysis on the features of each approach and also an overview of their positive and negative points. Finally, we apply a filter on those approaches to select the related works of our interest for our study and discuss their characteristics.

This chapter is organized as follows: Section 4.1 presents the proposed categorization; Section 4.2 applies the categorization on the approaches identified in the literature review; Section 4.3 defines the scope of our research and presents the related works; and Section 4.4 presents the preliminary considerations of this chapter.

4.1 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 similar to each other (JACOB, 2004), 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, 2016). They are: *(i)* the **Data Source**: classifies the type and quality of the input data to be used; *(ii)* the **Graphic Representation**: establishes characteristics and rules adopted in the representation of the concept map.

The categories for each perspective, respectively, are presented in Figure 4.1. 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.

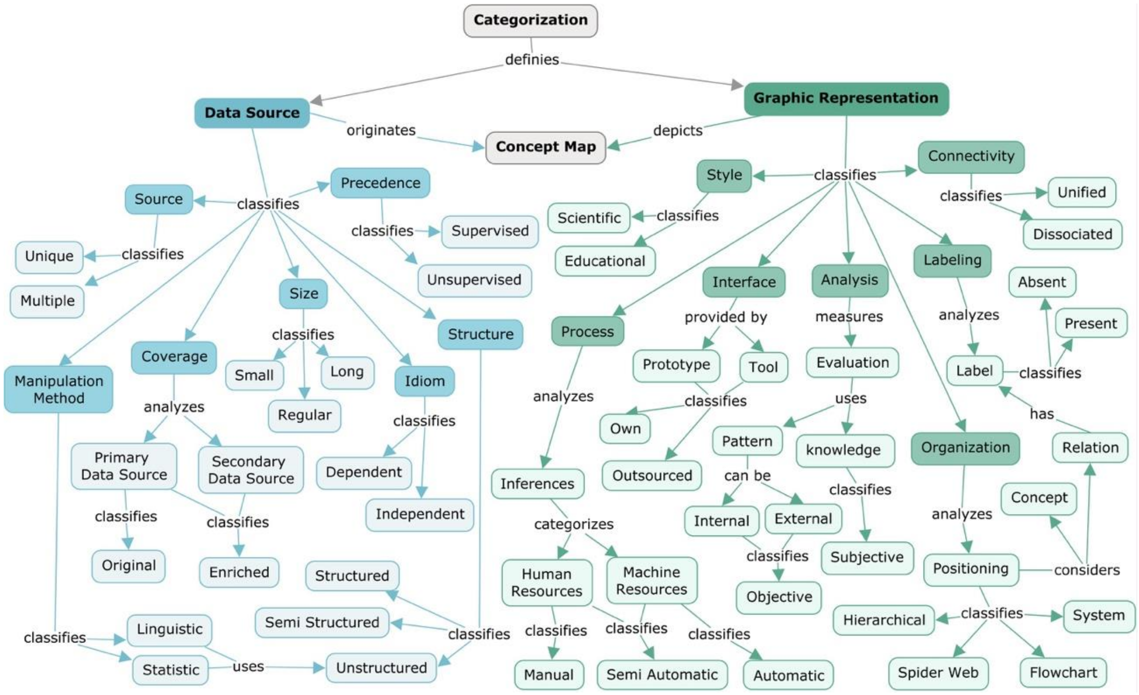


Figure 4.1 Concept map containing the perspectives and categories defined

4.1.1 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 Figure 4.1.

The category **Structure** analyzes the logical structure of how information is organized in the data source. It is classified as: *(i)* **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; *(ii)* **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 (LI, et al., 2008), OWL and RDF files, since RDF and OWL are documents encoded in XML; and *(iii)* **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: *(i)*

Linguistic: based on linguistic techniques (PÉREZ & VIEIRA, 2005), including, for example, linguistic pattern extraction, syntactic analysis, semantic analysis, context identification etc.; *(ii)* **Statistic**: based on calculations of statistical measures that detect new concepts and relationships (PÉREZ & VIEIRA, 2005), 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: *(i)* **Small**: text formed by some sentences, such as an abstract; *(ii)* **Regular**: text consisting of a few pages, such as an article, web page, didactic text and others; and *(iii)* **Long**: text consisting of many pages, such as a dissertation and thesis.

We can classify the **Source** category as: *(i)* **Unique**, when the use of only one data source is necessary and sufficient for the identification and extraction of the map elements; and *(ii)* **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 *(i)* **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 *(ii)* **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.

4.1.2 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 (VEKIRI, 2002). 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 Figure 4.1.

The **Analysis** identifies the type of devices used to evaluate the results. Thus, we classify the analysis as: *(i)* **Subjective**: when using the knowledge of user or a domain expert to assess the outcome; *(ii)* **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: *(i)* **Automatic**: when the intervention occurs only with machine resources from the choice of the data source to the construction of the concept map; *(ii)* **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 *(iii)* **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 (PÉREZ & VIEIRA, 2005), Graphviz (ZUBRINIC, et al., 2012) and WebDot (CHEN, et al., 2008).

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: *(i)* **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 *(ii)* **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: *(i)* **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 *(ii)* **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: *(i)* **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 *(ii)* **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 an 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 (TAVARES, 2007) we analyzed the **Organization** of the elements on the map generated by approaches according to: *(i)* **Hierarchical**: identified in most approaches, it organizes the concepts in order of importance, locating the more general at the top of the map; and *(ii)* **Spider web**: organizing the central and most important concept in the middle of the map; *(iii)* **Flowchart**: not identified in any of the studied approaches, organizes the concepts linearly including start and end points; and *(iv)* **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 (CHEN, et al., 2008), whose map organization, hierarchical or spider web depends on the purpose of the author.

4.2 Categorization applied to Literature Review

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. (PETERSEN, et al., 2008). The study consists of the following steps described respectively in sections 5.1, 5.2 and 6: *(i)* defining research questions, *(ii)* conducting research on primary studies, *(iii)* data extraction, and *(iv)* data analysis.

4.2.1 Research Questions

The initial question that motivated this review 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?

4.2.2 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 4.1 summarizes the selection process and presents the number of papers identified at each step.

Source	Studies	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*
ScienceDirect	*21*	*16*	*2*	*1*	*0*	*0*	*14*
Total	***134***	***55***	***27***	***5***	***1***	***0***	***30***

Table 4.1 Selection process of primary study

Although the search was not limited to a particular period, all studies were found between the years 2001 and 2016. The graph of Figure 4.2 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.

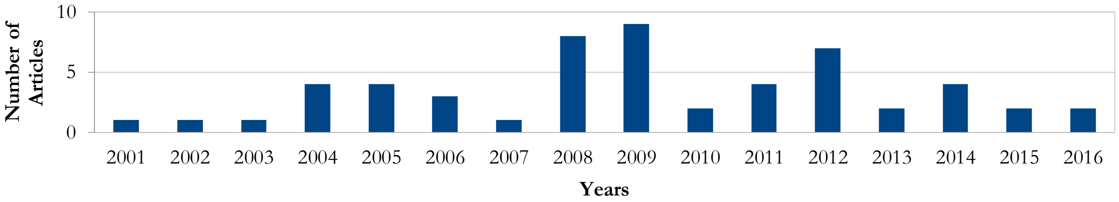


Figure 4.2 Concentration of studies per year

4.2.3 Analysis on the Categorization

To answer the research questions in Section 4.2.1, the categorization proposed in Section 4.1 was adopted as a metric for analyzing the technological approaches selected in the primary study. Table 4.2 synthesizes the result of the categorization performed for the 30 selected papers, divided into two main areas: (i) **Categories**: located on the left side, horizontally arranged, associates the Reference area with the categories of Data Source, and Graphic Representation; (ii) **References**: on the right side, arranged vertically, denotes the approach identified by its number in the list of references to follow:

> 1 (WANG, et al., 2008), 2 (TSENG, et al., 2007), 3 (PIPITONE, et al., 2014), 4 (QASIM, et al., 2013), 5 (WANG & LIU, 2016), 6 (RICHARDSON & FOX, 2005), 7 (OLNEY, et al., 2011), 8 (DE LA VILLA, et al., 2012), 9 (LEAKE, et al., 2004), 10 (ZUBRINIC, et al., 2012), 11 (AL-SAREM, et al., 2011), 12 (LIPIZZI, et al., 2016), 13 (BICHINDARITZ & AKKINENI, 2006), 14 (LEE & SEGEV, 2012), 15 (CHEN, et al., 2006), 16 (CHEN & SUE, 2013), 17 (LEE, et al., 2015), 18 (LEE, et al., 2009), 19 (BAI & CHEN, 2008), 20 (PIRNAY-DUMMER & IFENTHALER, 2011), 21 (AJLI & AFDEL, 2014), 22 (LEE, et

al., 2012), 23 (YI & LI, 2014), 24 (ELHOSEINY & ELGAMMAL, 2012), 25 (LAU, et al., 2009), 26 (BAI & CHEN, 2008), 27 (CHEN, et al., 2008), 28 (ŽUBRINIĆ, et al., 2015), 29 (KARANNAGODA, et al., 2013) e 30 (ZOUAQ & NKAMBOU, 2009).

To understand the data represented in the table, 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.

Categories			References																													
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*
DATA SOURCE	**Precedence**	*Supervised*		■	■				■	■	■						■	■	■	■	■		■	■	■	■		■	■	■		
		Unsupervised	■			■	■	■				■	■	■	■	■						■					■				■	■
	Idiom	*Dependent*	■			■		■	■	■		■		■	■							■				■	■			■	■	■
		Independent		■	■		■				■		■				■	■		■	■		■		■			■				
	Structure	*Structured*																														
		Unstructured	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
		SemiStructured																														
	Coverage	*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	■	■	■	■	■	■	■	■			■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Table 4.2 Categorization applied to the approaches of primary study

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 *Absent* labeling (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 analysing 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 signalled by the categories:

(i) **Precedence**: Most approaches adopt the category Supervised and hence they limit the construction of the map to a previously known domain.

(ii) **Purpose**: Approaches to constructing concept maps from texts have been developed with the purpose of evaluating learning and representing text.

(iii) **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.

(iv) **Interface**: Although most of the approaches adopt their own interface, they do not develop the interface potential for learning beyond the graphic representation.

(v) **Labeling**: 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.

(vi) **Connectivity**: although most approaches build unified maps, ensuring this feature is a challenge that in most approaches is related to the text.

4.2.3.1 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 Figure 4.3 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.

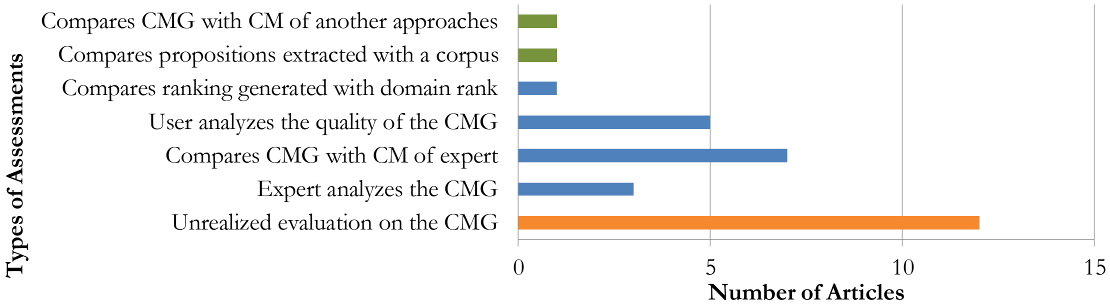


Figure 4.3 Type of assessments performed by the studies

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.

4.2.3.2 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:

(i) Stemming or lemmatization. Lemmatization is used to find the "lemma" of the word, disregarding grammatical changes such as tense and plurality (BIBER, et al., 1998). The main purpose of stemming is to reduce different grammatical forms to the "root" form.

(ii) Co-reference resolution, it is the task of finding all expressions that refer to the same entity in a discourse (LEE, et al., 2013). It is common that the candidates compete to be the antecedent of an anaphor (MITKOV, 2014).

(iii) 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 (NADEAU & SEKINE, 2007).

(iv) Stop words deletion as well as the removal of all information that does not constitute knowledge in the text;

(v) Multi-words and acronym identification;

(vi) 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:

(i) 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*".

(ii) 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).

(iii) 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 (FELDMAN & SANGER, 2007).

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

(i) Clustering is a descriptive task in which one seeks to identify a finite set of clusters to describe the data (KANTARDZIC, 2011) based on associating among features within the data, on the contexts they have in common (STRZALKOWSKI, 1999). Usually used to discover group of relevant concepts.

(ii) The frequency of terms is based on the assumption that the weight or relevance of a term occurring in a document is proportional to its frequency (LUHN, 1957).

(iii) Association rules are created by analyzing 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 4.3 synthesizes the main techniques identified.

The table is divided into two main areas: *(i)* **Techniques**: located on the left side, horizontally arranged, associates the Reference area with the techniques identified; *(ii)* **References**: on the right side, arranged vertically, denotes the approach identified by its number in the list of references presented in Table 4.2.

Techniques	References																													
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*
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								■					■																	

Table 4.3 Techniques identified in the Approaches

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

(i) Approaches adopting a linguistic method correspond to only 36% of the studied approaches, since this method requires more computational effort and have some limitations.

(ii) Normalization techniques are used both by linguistic and statistical methods.

(iii) Terms frequency techniques (60%) are primarily used by statistical methods, however they can be adopted in linguistic methods for the identification of relevant elements.

(iv) Some approaches use knowledge base (7%), ontologies (7%) or thesaurus (3%) as source for the identification of elements belonging to a domain.

4.3 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:

(i) Approach that uses only a single unstructured text of regular size;

(ii) Approach that does not use the knowledge of the user to modify the data source and accepts text from any domain;

(iii) Approach that adopts only machine resources;

(iv) Approach that generates maps containing labels for concepts and relationships. Moreover, the resulting map must not have fragmented portions;

(v) Approach that represents only the information contained in its own data source.

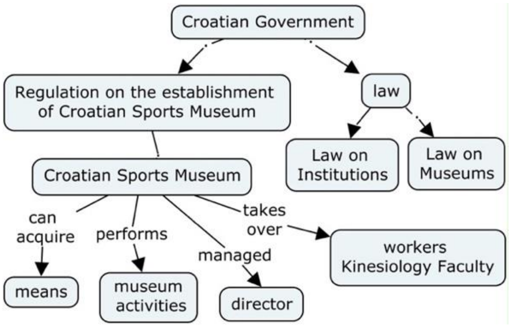
To fulfill these requirements, the following filters were applied to the categorization conducted on the literature review in Section 4.2:

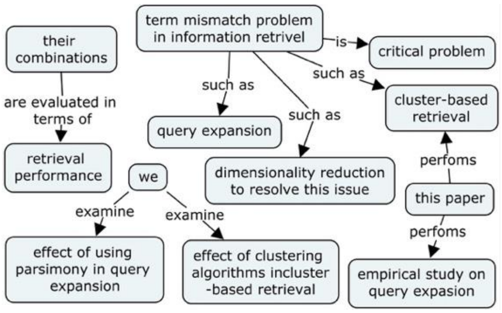
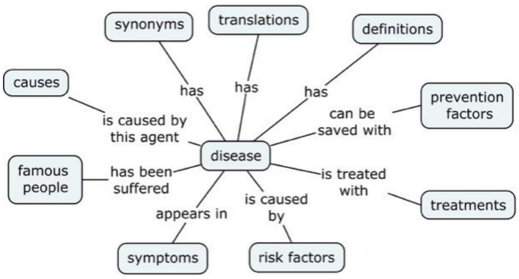
> Style (*Scientific*), Purpose (*Summarization*), Precedence (*Unsupervised*), Idiom (*all*), Structure (*Unstructured*), Coverage (*Original*), Source (*Unique*), Size (*regular or long*), Manipulation Method (*all*), Connectivity (*Unified*), Organization (*all*), Analysis (*all*), Interface (*all*),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

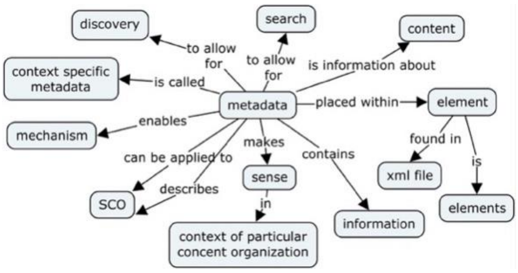
Since none of the approaches fulfilled the requirements of the mentioned filters, we kept only *Style*, *Structure* and *Labeling*. Accordingly, four approaches fulfilled these requirements and are synthesized in Table 4.4 where the right side provides an overview of the approaches and the left side 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 (WANG, et al., 2008); the approach assigns very long labels to concepts (WANG, et al., 2008); it accepts pronouns as labels for concepts (WANG, et al., 2008); it accepts prepositions as labels for relationships (WANG, et al., 2008); it accepts relationships without label (ŽUBRINIĆ, et al., 2015); map is created on a specific domain (DE LA VILLA, et al., 2012) (ŽUBRINIĆ, et al., 2015); the map is closer to representing the domain of the subject than the real content of the text (DE LA VILLA, et al., 2012); map is created using a set of domain documents (ZOUAQ & NKAMBOU, 2009) or a small text containing some sentences (DE LA VILLA, et al., 2012); it uses other domain data source in addition to the text such as ontology (ZOUAQ & NKAMBOU, 2009) (DE LA VILLA, et al., 2012), knowledge base (DE LA VILLA, et al., 2012) and thesaurus (ŽUBRINIĆ, et al., 2015); and does not show the direction (uses no arrow) of the association between concepts (DE LA VILLA, et al., 2012).

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

Approaches Detail	Concept Map
The approach (ŽUBRINIĆ, et al., 2015) 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.	 Croatian Government Regulation on the establishment of Croatian Sports Museum law Law on Institutions Law on Museums Croatian Sports Museum can acquire performs managed takes over means museum activities director workers Kinesiology Faculty Concept map generated by (ŽUBRINIĆ, et al., 2015) (translated)

<table>
<tr><th>Approaches Detail</th><th>Concept Map</th></tr>
<tr><td>The approach (WANG, et al., 2008) 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.</td><td>
their combinations
term mismatch problem in information retrivel
is
critical problem
such as
cluster-based retrieval
are evaluated in terms of
such as
query expansion
such as
perfoms
retrieval performance
we
dimensionality reduction to resolve this issue
this paper
examine
examine
perfoms
effect of using parsimony in query expansion
effect of clustering algorithms inclusters -based retrieval
empirical study on query expasion
Concept map generated by (WANG, et al., 2008)</td></tr>
<tr><td>The approach (DE LA VILLA, et al., 2012) 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.</td><td>

Concept map generated by (DE LA VILLA, et al., 2012)</td></tr>
<tr><td>The approach (ZOUAQ & NKAMBOU, 2009) 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.</td><td>

Concept map generated by (ZOUAQ & NKAMBOU, 2009) (clipping)</td></tr>
</table>

Table 4.4 Approaches identified from the Categorization

4.4 Some Considerations on the Chapter

This chapter presented a categorization of technological approaches for construction of concept maps and a literature review of their application between the years 1994 and 2015.

Since the categorization enabled a general and objective view of the approaches within the context of our interest, we filtered the literature review to select the related works which fulfilled the requirements of our research. Familiarizing ourselves with the related works, it was possible to analyze the process of construction adopted by each approach, as well as the concept maps built.

According to the analysis conducted in this section, it was possible to identify some challenges in the construction of concept maps. The next chapter will present the proposed conceptual model that aims to contribute to the solution of some of those challenges.

Chapter 5
The Conceptual Model

The objective of this research is to "develop a computational architecture to automatically build concept maps of scientific style as summarization of academic texts", as presented in Chapter 1. In this chapter we will present the proposal of a conceptual model that receives, as input, a text in English or Portuguese language, in pdf format, and returns, as output, an automatically constructed concept map. The constructed concept map can be of four distinct representation types: Text-based, Text-based Summarization, Text and Domain based Summarization, and Domain-based Summarization. The model is supported by the Concept Map Mining process already presented in Section 2.3.

This chapter is organized as follows: Section 5.1 describes the characterization adopted for conceptual model elaboration; Section 5.2 presents the overview of the conceptual model; Section 5.3 provides a detailed vision of this model and its activities; Section 5.4 presents the domain thesaurus model; Section 5.5 shows a service-oriented model for the communication and integration of the parts which compose the conceptual model; Section 5.6 presents some considerations on the chapter.

5.1 The Categorization

In the following, we present the characteristics identified and adopted for the elaboration of the conceptual model following the two perspectives presented by the proposed Categorization (see Chapter 4).

In order to be a comprehensive approach, the perspective **Data Source** uses an **Unstructured** text in the English or Portuguese language. The text is derived from academic articles in **Regular Size**, **Original Coverage** and **Unique Source**, since its goal is the representation of knowledge extracted from the input text itself.

The **Precedence** is classified as **Unsupervised** identifying the text domain by means of an automatic process. For this, the approach proposes the combined use of clustering and classification techniques that will identify the text domain.

In accordance with the **Manipulation Methods** which have been proposed and analyzed, we use a **Linguistic** method**.** Consequently, due to the selected manipulation method, the **Idiom** becomes **Dependent** and is defined to be in English or Portuguese.

From the perspective **Graphic Representation**, the approach adopts the **Automatic Process**, since it aims to build the concept map from the concepts and relationships extracted from the data source without any support from the user. If the representation is not built exactly as the user expects, he/she may use other tools to modify or enrich it. The constructed concept map will be displayed in its **Own Interface** with **Hierarchical Organization**.

The approach attaches a great deal of effort and importance to the categories **Present Labeling** and **Unified Connectivity**. Moreover, we adopted the **Scientific Style** with emphasis on knowledge engineering, since we want to rescue the fidelity of the concept map and enable the sharing and expansion of knowledge through other platform modules.

Finally, the map built will be **Analyzed Subjectively** as well as **Objectively**, using comparison with other concept maps and the analysis of the retrieved elements with other approaches.

5.2 Conceptual Model

In this section, we present an overview of the conceptual model designed to fulfill the objective proposed by this research. This model is supported by the conceptual and technological works discussed in Chapter 2 and Chapter 3, and by the literature review discussed in Chapter 4.

The conceptual model was developed on a Web environment that allows interaction of the user. It consists of the following components: **Domain Thesaurus**, responsible for storing data about the domains, and by **Formatter**, **Elements Extractor**, **Domain Identifier**, and **Summarizer** servers, responsible for processing the information. Figure 5.1 presents an overview of the model and its components.

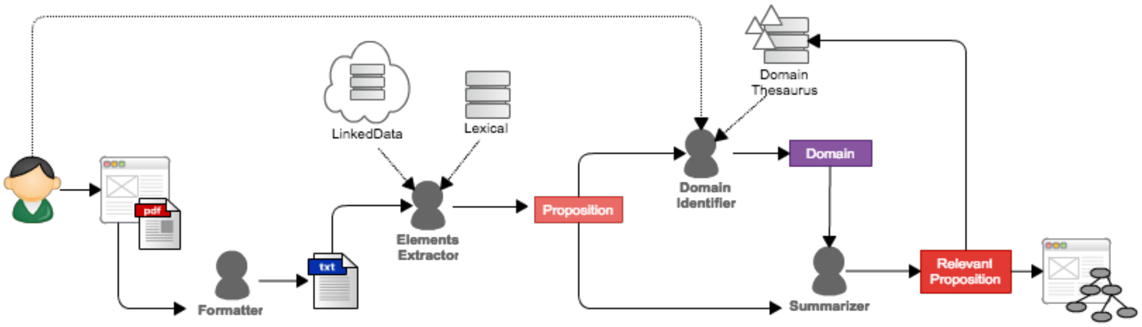


Figure 5.1 Conceptual model overview

A synthesis of the process shown in figure can be described as follows: the User accesses a web application and uploads a data source in pdf format. Then the **Formatter Server** turns that pdf into an unformatted text. From that text, the **Elements Extractor Server** extracts a set of propositions using a lexical dictionary and the Linked Open Data cloud for their labeling. Then, the propositions are used by the **Domain Identifier Server** to find a domain reference within the Domain Thesaurus. The propositions and, optionally, the concepts of the reference domain are used by the **Summarizer Server** to generate a concept map containing relevant propositions. Such propositions are returned to the Web application where the concept map is presented.

5.2.1 Domain Thesaurus

The thesaurus is responsible for storing data about domains. We modeled it as a graph structure composed by the nodes:

(i) *Concept*: One or more words extracted from the text that represents a meaning;

(ii) *Relation*: One or more words extracted from the text that represents an action;

(iii) Proposition: Composed of a Concept-Relation-Concept;

(iv) *Text*: Input text used to extract concepts and relations;

(v) *Cluster*: Collection of text containing similar characteristics;

(vi) *Class*: Label representing the text.

We define *domain* as a *cluster* that has one or more *classes*. Figure 5.2 shows the model of the Thesaurus and the relationship between its nodes.

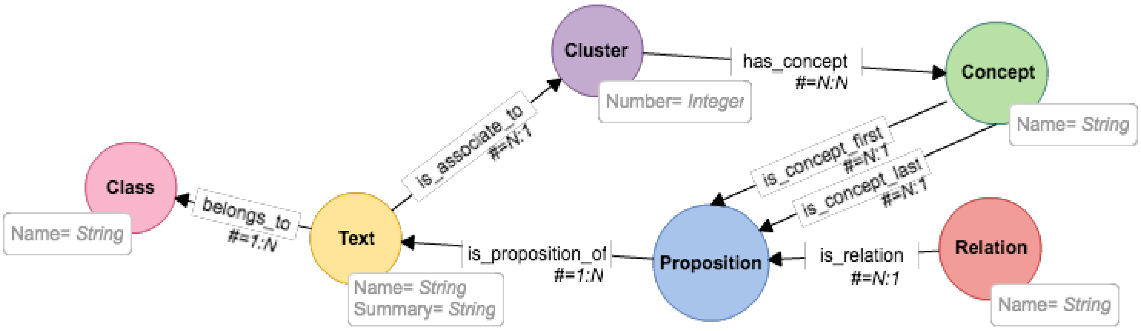


Figure 5.2 Model of the Domain Thesaurus

The thesaurus is used to store a conceptual vocabulary in a defined domain. The thesaurus acts in two stages:

Setter: adds or modifies the extracted elements from the text in the thesaurus.

Getter: uses the thesaurus to assist the identification of relevant elements of domain during the Ranking step.

In the beginning, the Thesaurus is empty, i.e., without any data about the domains. As the texts are being processed, the Thesaurus is filled with the data extracted from the input texts, forming a set of clusters with their respective concepts.

Consider, for example, the following text "*Concept map are graphical tool for organizing and representing knowledge*.". After the Setter stage the clipping shown in Figure 5.3 can be extracted from the thesaurus, where the colors purple, yellow, pink, blue, green and red respectively represent the nodes cluster, text, class, proposition, concept and relation.

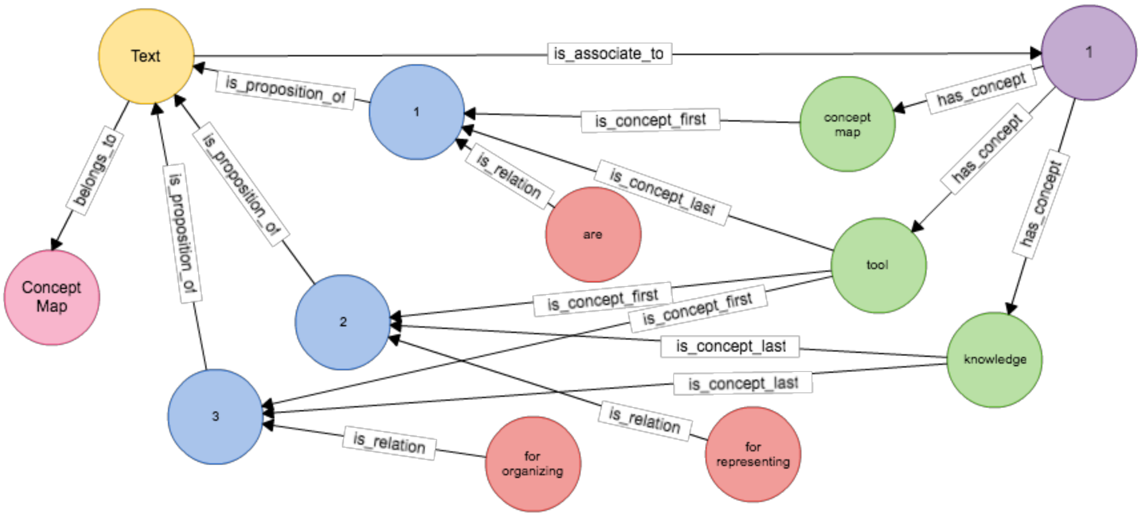


Figure 5.3 Example of the Domain Thesaurus

Therefore, we consider that Thesaurus is dynamic and evolves gradually in a knowledge base for different domains. This knowledge base will be represented by a future scheme for reuse and sharing by other projects.

5.2.2 Service-Oriented Model

Since the proposed conceptual model consists of four different servers and requires the interaction between them and with the user, we need to define a model that enables communication and integration of all parties. Aiming to provide a modularized web solution, accessible from anywhere, easily extended and embedded, we created a model based on services with fine granularity in order to provide flexibility and reuse of services.

In the following, we present the service-oriented model (Figure 5.4) that groups the services into four independent servers: *Formatter*, *Element Extractor*, *Domain Identifier*, and *Summarizer* Server. To complete the model, we have an application that recognizes the service interfaces, controls the interaction with the user, and manages the process.

The model is divided into three layers as follows:

(i) **Presentation**: the communication interface between service and client applications, in order to request the execution of a service and return its result;

(ii) **Service**: responsible for publishing services and communication with the data layer;

(iii) **Data**: stores the data, i.e., the thesaurus and the lexical dictionary.

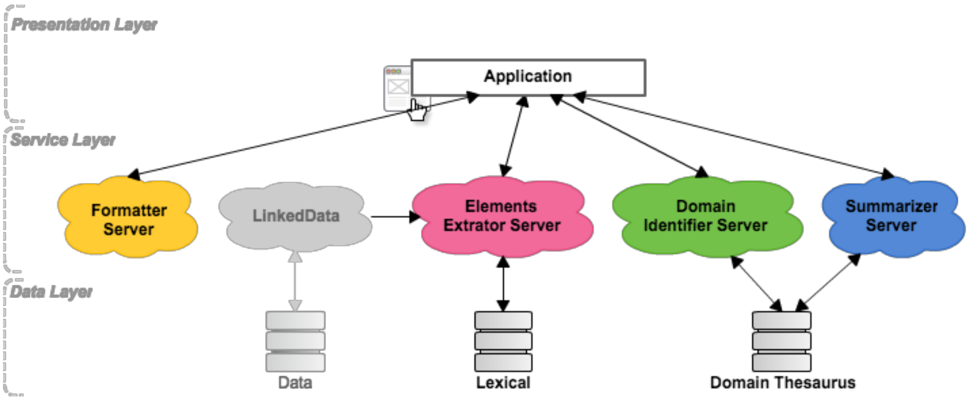


Figure 5.4 Service-Oriented model

5.2.2.1 Services Modeling

In order to allow any client application to be used on the modeled servers, Table 5.1 describes the service modeling.

Server	*Service*	*Description*
Formatter	GetFormatText	Returns formatted text according to the expected pattern.
Elements Extractor	GetExtractPropositions	Returns the list of propositions extracted from the text.
Domain Identifier	GetSimilarCluster	Returns the cluster into the thesaurus which is similar to the input text.
Domain Identifier	GetSimilarClass	Returns the class into the thesaurus which is similar to the input text.
Domain Identifier	GetDomainPropositions	Returns the propositions of a specific domain.
Domain Identifier	SavePropositionToThesaurus	Saves in thesaurus the relevant propositions list of input text.
Summarizer	GetRankingConcepts	Returns the list of ordered concepts according to their relevance.
Summarizer	GetRelevantPropositions	Identifies the propositions list which are relevant to the input text.

Table 5.1 Services of model

5.3 The Use Case Diagram

For the service-oriented model defined in the previous section, we present, in Figure 5.5, the use case diagram elicited.

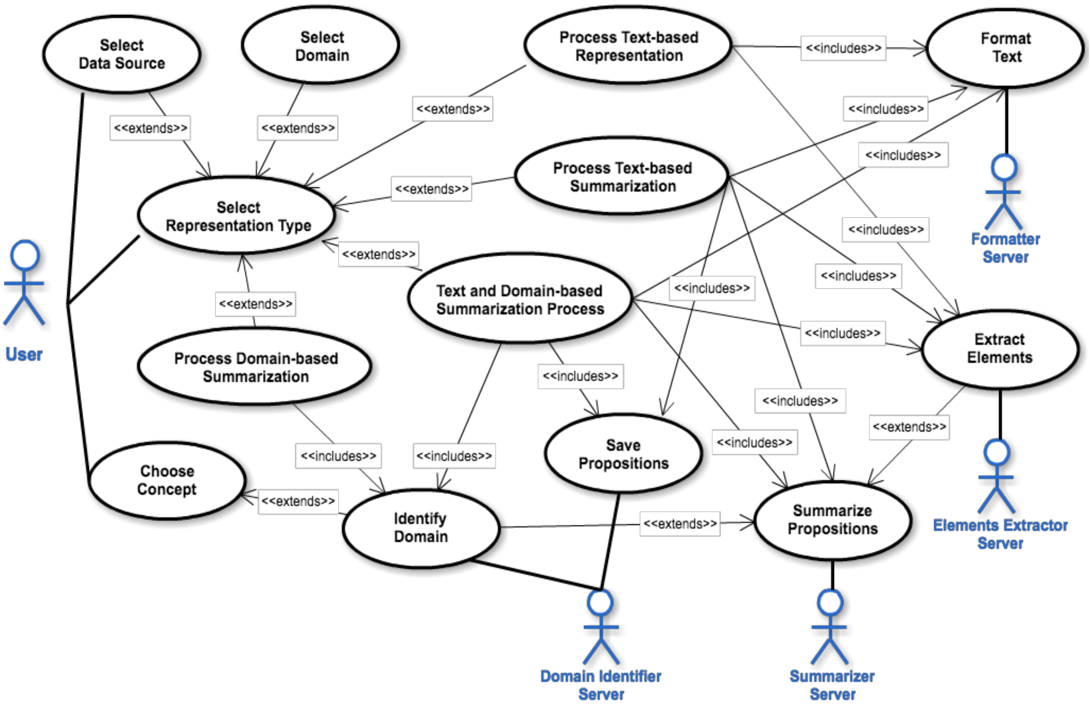


Figure 5.5 Use case diagram

Further, we present a brief description of the use cases:

Use Case I – Select Representation Type: The user's first action will be the selection of a representation type suitable for his/her purpose. It is possible to generate four different types of concept map representations:

(i) Text-based: Concept map containing *all possible* propositions extracted from the *text*;

(ii) Text-based Summarization: Concept map containing *exclusively the relevant* propositions extracted from the *text*;

(iii) Text and Domain based Summarization: Concept map containing *exclusively the relevant* propositions extracted from the *text* and thesaurus *domain*;

(iv) Domain-base Summarization: Concept map containing *exclusively the relevant* propositions extracted from the thesaurus *domain*;

Use Case II – Process Text-based Representation: this represents all possible propositions found in a text, i.e., the result of the activities performed by the Element Extractor Server;

Use Case III – Process Text-based Summarization: this represents exclusively the relevant propositions to a text, i.e., the result of the activities performed by the Summarizer Server on the result of Element Extractor Server.

Use Case IV – Text and Domain-based Summarization Process: this represents relevant propositions to the text with domain support, i.e., the result of the activities performed by the Summarizer Server on the result of Element Extractor and Domain Identifier Servers.

Use Case V – Process Domain-based Summarization: this represents exclusively the relevant propositions to a Domain, i.e., the result of the activities performed by the Summarizer Server on the result of Domain Identifier Server.

Use Case VI – Select Data Source: If the user selects the representation type Text-based Representation, Text-based Summarization or Text and Domain-based Summarization, he/she must select a data source in pdf format. This text will be the input source for the whole process.

Use Case VII – Select Domain: If the user selects the representation type Domain-based Summarization, the user must select a domain within the Thesaurus. The concepts of this domain will be the input source for the whole process.

Use Case VIII – Format Text: After selecting a text and a representation type, the Application, under the user's order, starts the sequence of activities to construction of the concept map from the text. For this, the Application requests the Formatter Server to clean the text and returns an unformatted text.

Use Case IX – Extract Elements: After formatting the text, the Application requests the Elements Extractor Server a proposition list extracted from the text.

Use Case X – Identify Domain: From the elements extracted by Use Case IX, the Application, with user´s help, requests the Domain Identifier Server the domain of that text.

Use Case XI – Choose Concept: If the domain of the text is not automatically identified during the Use Case X, the user chooses a concept to represent that text.

Use Case XII – Summarize Propositions: From the elements extracted by Use Case IX, the Application may request the list of relevant propositions from Summarizer Server.

Use Case XIII – Save Propositions: The propositions extracted from Use Case III and IV are stored in the Domain Thesaurus.

5.4 The Sequence Diagram

In the following, we present the sequence diagram that models the interaction between servers during the main processes. The exchange of messages is performed by requests and responses to and from the services.

5.4.1 Processing Text-based Representation

Figure 5.6 shows the sequence diagram containing the interactions between the components during the Use Case: Processing Text-based Representation.

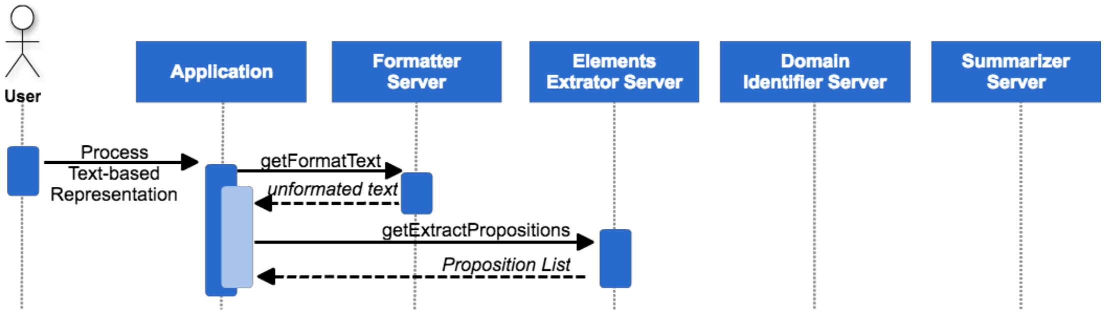


Figure 5.6 Sequence diagram of the use case: Process Text-based Representation

5.4.2 Process Text-based Summarization

Figure 5.7 shows the sequence diagram containing the interactions between the components during the Use Case: Process Text-based Summarization.

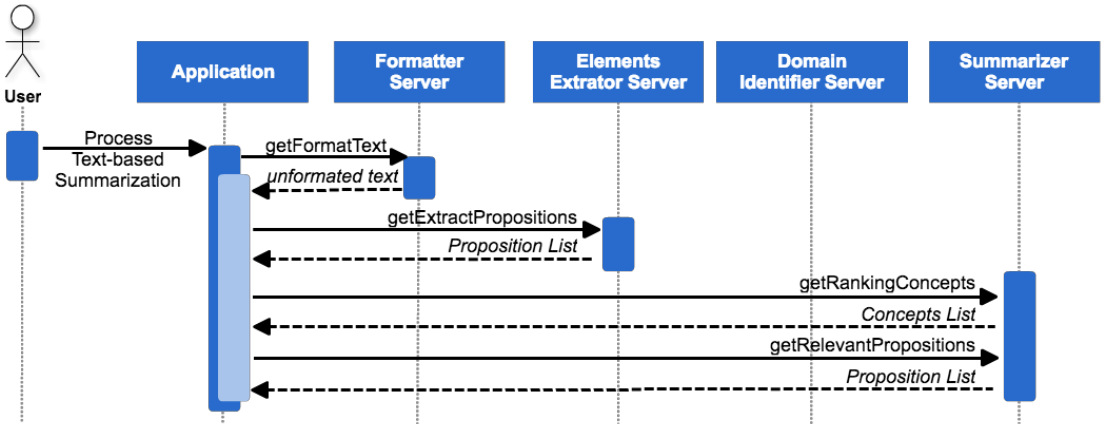


Figure 5.7 Sequence diagram of the case use: Process Text-based Summarization

5.4.3 Text and Domain-based Summarization Process

Figure 5.8 shows the sequence diagram containing the interactions between the components during the Use Case: Text and Domain-based Summarization Process.

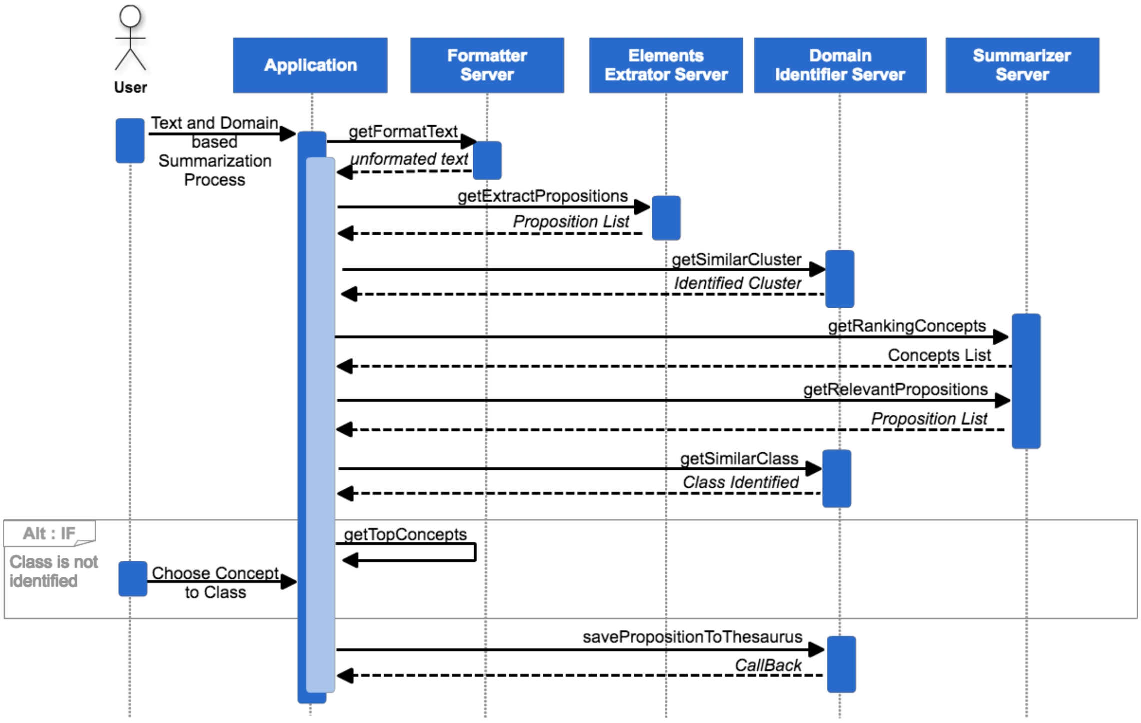


Figure 5.8 Sequence diagram of use case: Text and Domain-based Summarization Process

5.4.4 Process Domain-based Summarization

Figure 5.9 shows the sequence diagram containing the interactions between the components during the Use Case: Process Domain-based Summarization.

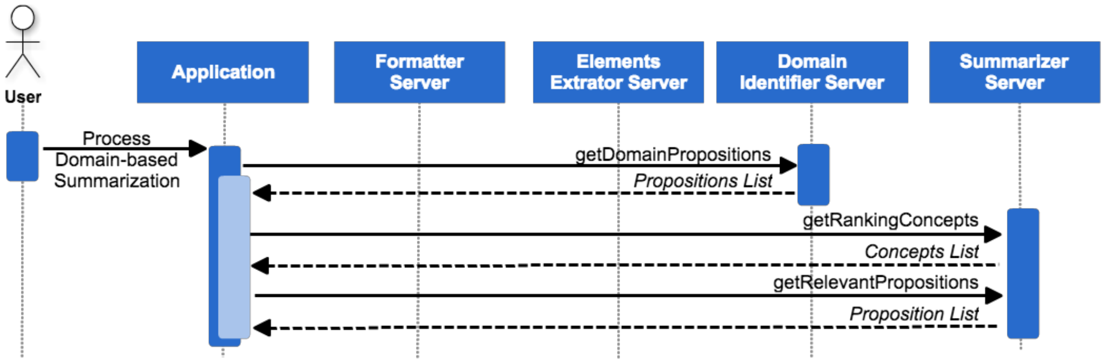


Figure 5.9 Sequence diagram of use case: Process Domain-based Summarization

5.5 Some Considerations on the Chapter

This chapter has proposed a conceptual model of a technological approach for the construction of concept maps. This model satisfies the premise of not being limited to a previously defined domain. Consequently, the model did not rely on the adoption of an ontology or a set of domain-specific texts for the information extraction. Instead, we developed a strategy based on a dinamic domain thesaurus.

We defined the thesaurus as knowledge base continuously enriched with data extracted from texts. As a consequence, the Thesaurus can be used as a data source for the construction of concept maps representing a whole domain, similar to a lightweight ontology. Moreover, the thesaurus might be used as a publicly available data source.

We also presented a service-oriented model for communication and integration of the constituent parts of the conceptual model. We adopted a fine granularity for the services ensuring greater flexibility and the reuse of services for other projects.

Finally, we emphasize that this conceptual model produces three important components: a concept map construction tool, a domain Thesaurus and an information extraction library.

The next chapter presents the technological architecture resulting from the models presented in this chapter.

Chapter 6
The Technological Architecture

This chapter presents a technological architecture based on the conceptual model proposed in Chapter 5. This architecture was designed for the Web scope.

The chapter is organized as follows: Section 6.1 describes scope of the architecture; Section 6.2 presents the detailed view of the technological architecture and its components; Sections 6.3, 6.4, and 6.5 describe the formatter, the elements extractor, the domain identifier, and the summarizer modules, respectively; Section 6.6 shows technical details about the service-oriented model; and in Section 6.7 we discuss some considerations on the chapter.

6.1 Scope of Technological Architecture

The architecture is designed for a web scope, since its product is a public tool accessible anywhere through an internet connection. In addition to online access, the architecture will offer services to be consumed or embedded in other projects.

Based on these premises, we define some requirements for the elicitation of the technical components which integrate the architecture:

(i) Web programming language with support for service and active development in NLP and IR solutions;

(ii) Libraries, resources and services with documentation and public access;

(iii) Current and minimally satisfactory technologies;

(iv) Prior knowledge about the technologies;

Therefore, the architecture was elaborated in Java language along with the Spring MVC framework.

6.2 Detailed View of the Technological Architecture

Based on the requirements elicited in Section 6.1, Figure 6.1 presents the technological architecture defined from the conceptual model proposed in Chapter 5. Bold letters such as **S**, **E**, **T** etc. is a stem for the technology used.

(S) – Stanford CoreNLP: an extensible pipeline that provides core natural language analysis such as Tokenization, Sentence Splitting, Part-of-speech Tagging, Morphological Analysis, Named Entity Recognition, Syntactic Parsing and Co-reference Resolution (MANNING, et al., 2014). The toolkit works with models in different languages such as

English. It is licensed under the GNU General Public License for download from http://stanfordnlp.github.io/CoreNLP.

(O) – Apache OpenNLP: is a machine learning based toolkit for natural language processing, able to perform Tokenization, Sentence Segmentation and Part-of-Speech Tagging. The toolkit works with models in different languages such as Portuguese. It is available as an Apache License 2.0 for download from https://opennlp.apache.org.

(T) – Apache Tika: a content analysis toolkit able to help identify the language of a piece of text, i.e., language detection (MATTMANN & ZITTING, 2011). It is available as an Apache License for download from https://tika.apache.org.

(W) – WS4J: a Java API for several published Semantic Relatedness/Similarity algorithms such as LIN. It is a re-implementation of WordNet-Similarity (PEDERSEN, et al., 2004). It is available as a GNU General Public License for download from https://code.google.com/archive/p/ws4j.

(A) – JawJaw: is a Java API that contains Princeton's English WordNet v3.0. It offers access to lexical knowledge of a given word such as hypernym, hyponym and definition in English. It is available as a General Public License for download from https://code.google.com/archive/p/jawjaw.

(J) – Apache Jena: a Java API which supports the creation, manipulation, and query of RDF graphs (MCBRIDE, 2001). It is available as an Apache License for download from www.hpl.hp.com/semweb/jena-top.html.

(K) – DKPro Similarity: an open source framework that offers a comprehensive repository of text similarity measures such as Cosine similarity (BÄR, et al., 2013). It is available as an Apache Software License for download from https://dkpro.github.io/dkpro-similarity.

(E) – ExtroutNLP: a Java API developed during this work that provides a suite of text processing libraries such as OpenIE, Ranking and Summarization for Portuguese and English languages. It is available as a GNU General Public License for download from http://extroutnlp.lied.inf.ufes.br.

WordNet: a large lexical database of English language (MILLER, 2005). Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), which are interlinked by means of conceptual-semantic and lexical relations. It is publicly available for download from https://wordnet.princeton.edu.

DBPedia: is served as Linked Data on the Web (AUER, et al., 2007). It is a source of knowledge by extracting structured information from Wikipedia and by making this information accessible on the Web. It is available under the terms of the Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License from http://dbpedia.org/snorql.

Neo4J: NoSql graph database with ACID features (VUKOTIC, et al., 2015). It is implemented in Java and uses the Cypher Query Language through a transactional HTTP endpoint. It is available in a GPLv3-licensed open-source and AGPLv3 Affero General Public License for download from https://neo4j.com.

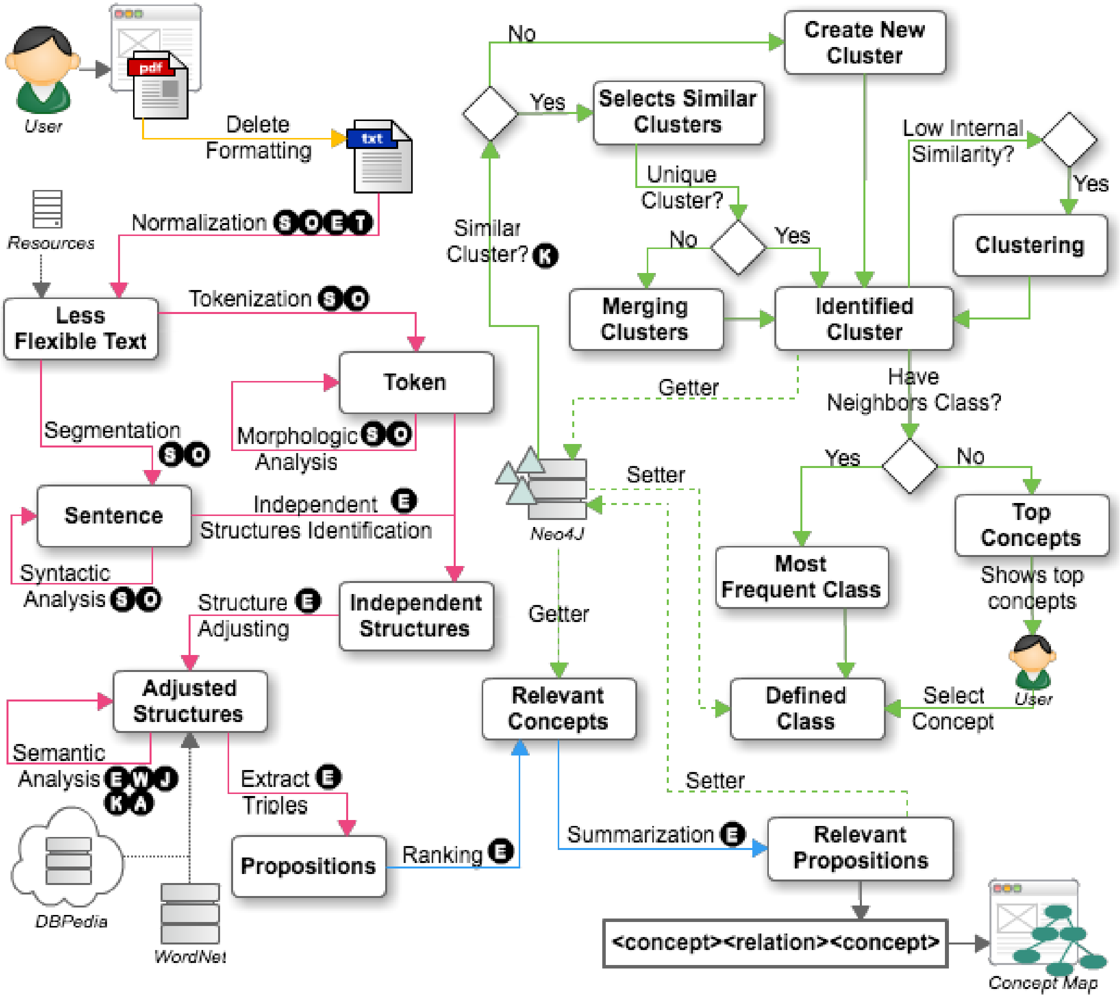


Figure 6.1 The Technological Architecture with the technologies in bold letters

The following sections describe in detail the technological components used in each module.

6.3 Element Extractor Module

It begins with the Normalization step and ends with the Concept Mapping step. The following sentences A (Figure 6.2) and B (Figure 6.3) are used as examples to the module.

> *"Concept maps were developed in 1972 in the course of Novak's research program at Cornell where he sought to follow and understand changes in children's knowledge of science. During the course of this study the researchers interviewed many children, and they found it difficult to identify specific changes in the children's understanding of science concepts by examination of interview transcripts. This program was based on the learning psychology of David Ausubel. The fundamental idea in Ausubel's cognitive psychology is that learning takes place by the assimilation of new concepts and propositions into existing concept and propositional frameworks held by the learner."*

Figure 6.2 Sentence A

> *"Os mapas conceituais foram desenvolvidos em 1972, dentro do programa de pesquisa realizado por Novak na Universidade de Cornell, no qual ele buscou acompanhar e entender as mudanças na maneira como as crianças compreendiam a ciência. Ao longo desse estudo, os pesquisadores entrevistaram um grande número de crianças e tiveram dificuldade em identificar mudanças específicas na compreensão de conceitos científicos por parte delas apenas examinando entrevistas transcritas. Esse programa se baseava na psicologia da aprendizagem de David Ausubel. A ideia fundamental na psicologia cognitiva de Ausubel é que a aprendizagem se dá por meio da assimilação de novos conceitos e proposições dentro de conceitos preexistentes e sistemas proposicionais já possuídos pelo aprendiz."*

Figure 6.3 Sentence B

6.3.1 Normalization

The Normalization step is responsible for making text clearer and cleaner. This is done using Stanford CoreNLP and ExtroutNLP through the following tasks:

Removing non-propositional sentence: A proposition is the true or false content expressed by an affirmation. Therefore, this step attempts to remove phrases that express orders, questions, or advice. For this we identify the last token of sentence using CoreNLP modules of Sentence Splitting, Tokenization and Part-of-Speech. If the token is type "." and contains the character "?" or "!", the sentence is removed.

Resolving Anaphora: To solve some anaphora types we use the Co-reference Resolution (LEE, et al., 2013) module of CoreNLP, a model that performs entity-centric co-reference, where all mentions that point to the same real-world entity are jointly modeled, in a rich feature space using solely simple, deterministic rules. Figure 6.4 shows one of the mentions identified by Co-reference Resolution in Sentence A.

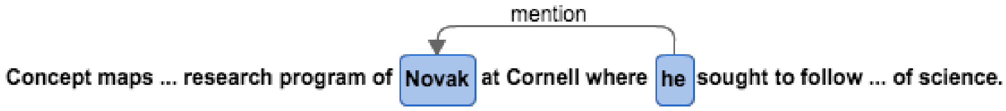

mention
Concept maps ... research program of
Novak
at Cornell where
he
sought to follow ... of science.

Figure 6.4 A mention identified in Sentence A

To resolve the anaphora, we seek all mentions of the type Proper noun or Nominal and replace them by their co-references.

Interpreting Genitive Case: The English language has two genitive constructions, the proposed, such as "John's book", and the postposed, such as "a book of John's" (LYONS, 1986). We chose not to represent genitive constructions in a proposition because it represents a very specific and extensive concept. Instead, we transform the genitive constructions into an intermediate representation composed by "*of*", such as "*book of John*". This representation is not the most appropriate for the English language, but it is satisfactory to represent propositions. For this, we use the GenitiveInterpretation module of API ExtroutNLP. It uses a parser dependency tree to identify genitive constructions and creates an intermediate representation with the preposition "*of*".

The following is the result of the Normalization step for Sentence A.

> *"Concept maps were developed in 1972 in the course of research program of Novak at Cornell where Novak sought to follow and understand changes in knowledge of children and science. During the course of this study the researchers interviewed many children, and researchers found the course of this study difficult to identify specific changes in the understanding of children and science concepts by examination of interview transcripts. This program was based on the learning psychology of David Ausubel. The fundamental idea in cognitive psychology of Ausubel is that learning takes place by the assimilation of new concepts and propositions into existing concept and propositional frameworks held by the learner."*

Figure 6.5 Normalization step for Sentence A

6.3.2 Tokenization and Morphological Analysis

The steps of tokenization and morphological analysis are performed on the normalized text. Tokens not included in Table 3.1 (Section 3.2.1) will remain in the sentence structure but will not be considered.

For the **English language**, we use the Stanford CoreNLP toolkit. The Tokenization is performed by means of Tokenization module with *PTBTokenizer* model based on Penn Treebank and provided by the toolkit. The Morphological Analysis is performed by means of Parts-of-Speech module with *left3words* model provided by the toolkit. Figure 6.6 shows the tokens and the morphological analysis for a part of Sentence A, i.e., the result of this step for English language.

> *Concept[NN] maps[NNS] were[VBD] developed[VBN] in[IN] 1972[CD] in[IN] the[DT] course[NN] of[IN] research[NN] program[NN] of[IN] Novak[NNP] at[IN] Cornell[NNP] where[WRB] Novak[NNP] sought[VBD] to[TO] follow[VB] and[CC] understand[VB] changes[NNS] in[IN] knowledge[NN] of[IN] children[NNS] and[CC] science[NN].*

Figure 6.6 Tokenization and Morphological Analysis for Sentence A

For the **Portuguese language**, we use Apache OpenNLP toolkit. The Tokenization is performed by means of Tokenizer module with *pt-token* model trained with CoNLL-X bosque corpus, provided by the toolkit. The Morphological Analysis is performed by means of PosTagger module with *pt-tagger-macmorpho* model trained with MacMorpho corpus and provided by the API ExtroutNLP. MacMorpho is a Brazilian texts corpus annotated with part-of-speech tags (FONSECA & ROSA, 2013). Both corpora, bosque and macmorpho, cannot be combined to provide a larger resource, since each one defines a different tagset. Therefore, a conversion of the tagset is performed. Figure 6.7 shows the tokens and the morphological analysis for a part of Sentence B, i.e., the result of this step for Portuguese language.

```
Os[ART] mapas[N] conceituais[ADJ] foram[V] desenvolvidos[PCP] em[PREP] 1972[N]
,[PU] dentro[PREP] do[PREP+ART] programa[N] de[PREP] pesquisa[N] realizado[PCP]
por[PREP] Novak[NPROP] na[PREP+ART] Universidade[NPROP] de[NPROP]
Cornell[NPROP] ,[PU] no[PRO-KS] qual[PRO-KS] Novak[NPROP] buscou[V]
acompanhar[V] e[KC] entender[V] as[ART] mudanças[N] na[PREP+ART] maneira[N]
como[PREP] as[ART] crianças[N] compreendiam[V] a[ART] ciência[N] .[PU]
```

Figure 6.7 Tokenization and Morphological Analysis for Sentence B

6.3.3 Text Segmentation and Syntactic Analysis

The normalized and tokenized text is divided into individual sentences for the syntactic analysis. The syntagms not included in Table 3.2 (Section 3.2.2) remain in the structure but are not considered.

For the **English language**, we use Stanford CoreNLP toolkit. The Text Segmentation is performed by means of Sentence Splitting module, a deterministic consequence of tokenization when a sentence-ending character (., !, or ?) is found which is not grouped with other characters into a token. The syntactic analysis is performed by means of Constituency Parsing module (KLEIN & MANNING, 2003) with probabilistic context-free grammars model provided by the toolkit.

For **Portuguese language**, we use Apache OpenNLP and Stanford CoreNLP toolkit. The Text Segmentation is performed by means of SentenceDetector module of OpenNLP with *pt-sent* model trained with CoNLL-X bosque corpus and provided by the toolkit. The Syntactic Analysis is performed by means of LexicalizedParser module of CoreNLP with *pt-parser-cintil* model trained with CINTIL Treebank corpus and provided by the API ExtroutNLP. CINTIL Treebank is a Portuguese corpus annotated with the representation of constituency relations (BRANCO, et al., 2010).

Figure 6.8 shows the Text Segmentation and Syntactic Analysis for a part of Sentence A, i.e., the result of this step for the English language.

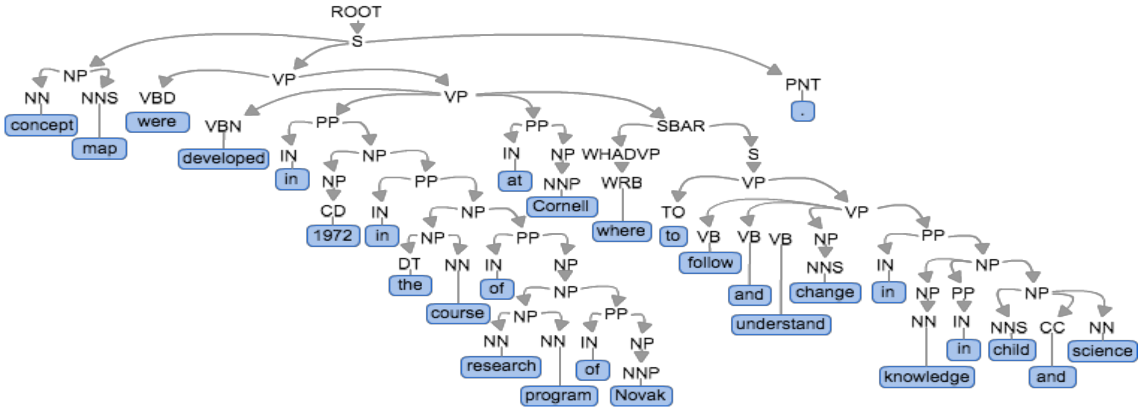


Figure 6.8 Text Segmentation and Syntactic Analysis for Sentence A

Figure 6.9 shows the Text Segmentation and Syntactic Analysis for a part of Sentence B, i.e., the result of this step for the Portuguese language.

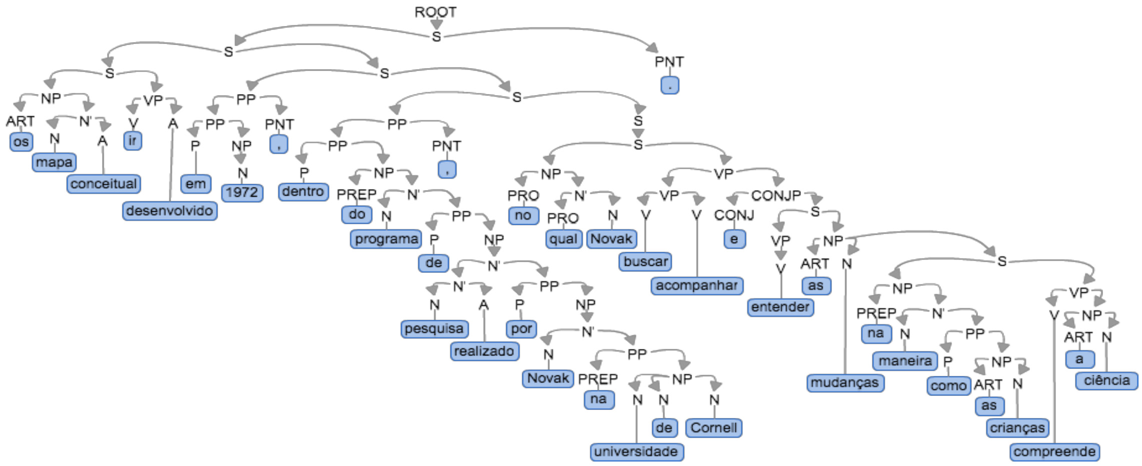


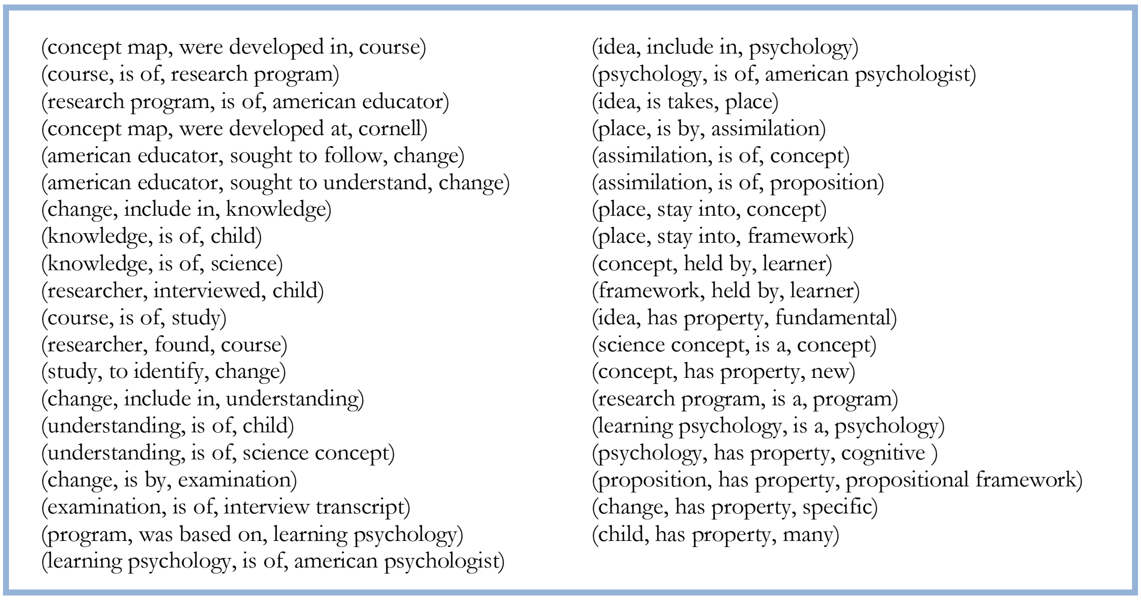
Figure 6.9 Text Segmentation and Syntactic Analysis for Sentence B

6.3.4 Extract Triples

To extract triples from the parser tree, we use the OpenIE module of API ExtroutNLP, which adopts deep search and heuristic rules.

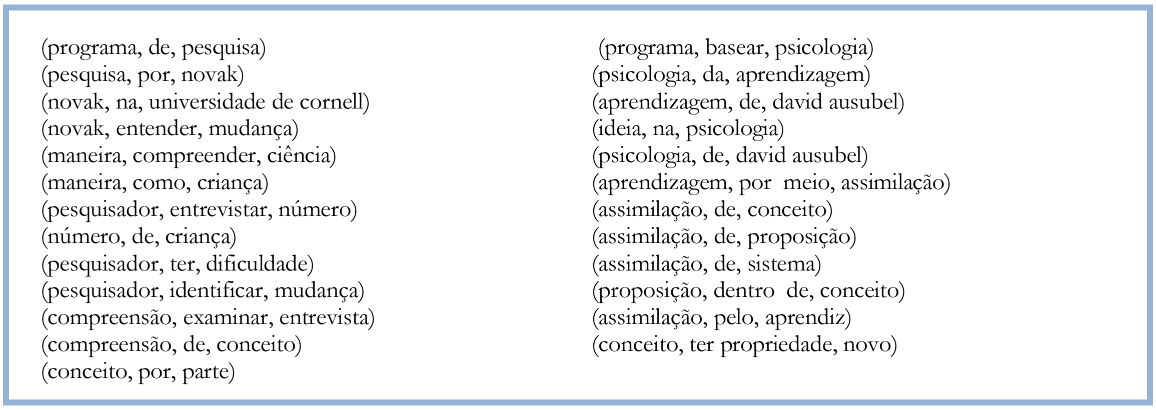
Each parser tree is segmented into a set of complete independent structures. These independent structures are adjusted according to the name similarities, relationship identification, lemmatization, and named entity interpretation in order to make the most concise and unambiguous structures. Finally, propositions in the form (*concept1, relation, concept2*) are extracted from the adjusted structures. This process is explained in Section 7.3.

Based on the parsing tree of Portuguese (Figure 6.8) and English (Figure 6.9), the following presents the propositions extracted from Sentences A (**Figure 6.10**) and B (**Figure 6.11**), i.e., the result of this step for the English and Portuguese languages.



(concept map, were developed in, course)
(course, is of, research program)
(research program, is of, american educator)
(concept map, were developed at, cornell)
(american educator, sought to follow, change)
(american educator, sought to understand, change)
(change, include in, knowledge)
(knowledge, is of, child)
(knowledge, is of, science)
(researcher, interviewed, child)
(course, is of, study)
(researcher, found, course)
(study, to identify, change)
(change, include in, understanding)
(understanding, is of, child)
(understanding, is of, science concept)
(change, is by, examination)
(examination, is of, interview transcript)
(program, was based on, learning psychology)
(learning psychology, is of, american psychologist)
(idea, include in, psychology)
(psychology, is of, american psychologist)
(idea, is takes, place)
(place, is by, assimilation)
(assimilation, is of, concept)
(assimilation, is of, proposition)
(place, stay into, concept)
(place, stay into, framework)
(concept, held by, learner)
(framework, held by, learner)
(idea, has property, fundamental)
(science concept, is a, concept)
(concept, has property, new)
(research program, is a, program)
(learning psychology, is a, psychology)
(psychology, has property, cognitive)
(proposition, has property, propositional framework)
(change, has property, specific)
(child, has property, many)

Figure 6.10 Extract Triples for Sentence A



(programa, de, pesquisa)
(pesquisa, por, novak)
(novak, na, universidade de cornell)
(novak, entender, mudança)
(maneira, compreender, ciência)
(maneira, como, criança)
(pesquisador, entrevistar, número)
(número, de, criança)
(pesquisador, ter, dificuldade)
(pesquisador, identificar, mudança)
(compreensão, examinar, entrevista)
(compreensão, de, conceito)
(conceito, por, parte)
(programa, basear, psicologia)
(psicologia, da, aprendizagem)
(aprendizagem, de, david ausubel)
(ideia, na, psicologia)
(psicologia, de, david ausubel)
(aprendizagem, por meio, assimilação)
(assimilação, de, conceito)
(assimilação, de, proposição)
(assimilação, de, sistema)
(proposição, dentro de, conceito)
(assimilação, pelo, aprendiz)
(conceito, ter propriedade, novo)

Figure 6.11 Extract Triples for Sentence B

6.4 Domain Identifier Module

The module adopts a thesaurus to store domain information and a supervised model to reduce text classification efforts. It receives, as input, a list of propositions and identify the domain from the Thesaurus.

Initially, the techniques proposed for the domain identification are not be very effective. However, as new texts are processed, the clusters are better identified and the thesaurus gradually expands and results in a kind of light ontology to each domain.

The purpose of this module is to decrease efforts to classify the new texts in a given domain within the thesaurus. An automatic text classification typically uses a large training set of labeled text at hand. As our approach is not limited to a single specific domain, we propose the use of a clustering process before the semi-automatic classification, based on the proposal of Oliveira (DE OLIVEIRA, 2015).

Thus, we define the model in two steps: *clustering*, to group texts of the same domain, and supervised *classification*, to assign a class to text in the domain, based on the labels assigned to thesaurus texts by previous users.

6.4.1 Cluster Identification

In the thesaurus, the set of concepts of each text belonging to the same domain is automatically defined as a *cluster*. Since the cluster gathers domain concepts as a whole, i.e., a large amount of concepts, we define a centroid to represent each cluster. The centroid is represented only by 60% of frequent concepts of each cluster.

This step receives, as input, a list of propositions which are mapped into a list of concepts. The relations are not used to identify the cluster. Each centroid and the mapped concepts list are transformed into vector space representation composed by the concepts. The clustering process is performed gradually through agglomerations and divisions of the clusters.

The first step is to calculate the cosine similarity between the mapped concepts list and each cluster centroid, using the framework DKPro. This step can result in the following actions:

(i) If the clusters similarity is less than 0.6, then a new cluster will be created.

(ii) If the clusters similarity is greater than or equal 0.6, then the mapped concepts list will be assigned to the cluster with the highest similarity, we adopt two different actions:

- For a single cluster, the mapped concepts list will be assigned to this cluster.
- If there are more than one cluster, the clusters will be merged to a single cluster.

At this point, an initial cluster is selected and its internal similarity is verified. The concepts of each text within initial cluster are represented in a vector space and the cosine similarity with its centroid is calculated. If the internal similarity is less than 0.75, the initial cluster is partitioned into two clusters.

Partitioning is performed using the k-means algorithm, Section 3.1.2.1, where k=2 and the two initial centroids are the concepts of the two most dissimilar texts within of cluster. The next steps follow the rules explained before, centroid is represented by frequent concepts in 60% of the texts of the cluster, and the texts are assign to their closest cluster using cosine similarity. This partitioning is repeated until the clusters have internal similarity greater than 0.75. At this point, the cluster with the highest similarity is identified.

The choice of k-means approach can be justified by the fact that these clusters are concise and have similar behavior. To satisfy the k value specification, we define it as a constant 2.

6.4.2 Class Identification

Once the cluster has been identified, a class must be assigned or chosen to represent the text, i.e., the mapped concepts list. The cluster is an unlabeled domain that contains a set of text with labeled class.

We use the KNN approach with cosine similarity to define the three closest neighbors of the text within the cluster, i.e., k = 3. This step may result in the following actions:

(i) If there are three nearest neighbors, the most frequent class among them will be assigned to the text.

(ii) Otherwise, top 6 concepts of the mapped concepts list will be suggested for the user as a new class. The user selects the concept that best represents the text, whereupon a new class will be defined and assigned to the text.

At the end of cluster and class identification steps, the text will belong to a class and will be associated to a cluster. Each cluster will result in a light ontology of a specific domain.

6.5 Summarizer Module

The module receives, as input, a list of propositions and returns, as output, the relevant propositions list.

As this module is independent of linguistic analysis, we will use only a text in the English language for demonstration. For this we will adopt the Sentence A (Figure 6.2) and the propositions list (Figure 6.10).

6.5.1 Ranking

This step is responsible for ordering the mapped concepts list, assigning a weight for each concept according to a metric. For this, we use the *Ranking* module of ExtroutNLP API, which adopts the HAF model. The HAF model is based on a graph representation and calculates the weight of vertices according to their input and output connections added to the frequency in the text and in the domain. This model is explained in Section 7.4.

For the frequency in the text, we use the frequency of the concept in the input text, and in the domain, we use the inverse frequency of the concept in the Thesaurus (Section 3.1.1.6). The weight calculated by the metric for each concept is assigned to the corresponding vertex (concept) in the graph. According to their weight, the concepts are ranked.

Based on the propositions list (Figure 6.10) and considering an empty thesaurus, the following (Figure 6.12) presents the ranking of the concepts identified.

Ranking	Weight	Concepts	Ranking	Weight	Concepts
1	0.8333	change	17	0.3176	concept map
2	0.6814	concept	18	0.2912	science concept
3	0.6490	course	19	0.2912	program
4	0.6226	child	20	0.2912	examination
5	0.5421	place	21	0.2912	proposition
6	0.5157	psychology	22	0.2647	learner
7	0.4431	idea	23	0.2245	framework
8	0.4167	research program	24	0.1657	fundamental
9	0.4167	understanding	25	0.1657	organization - cornell
10	0.4167	knowledge	26	0.1657	new
11	0.4167	assimilation	27	0.1657	many
12	0.4167	learning psychology	28	0.1657	specific
13	0.3843	researcher	29	0.1657	interview transcript
14	0.3579	study	30	0.1657	cognitive
15	0.3579	american educator	31	0.1657	science
16	0.3314	american psychologist	32	0.0990	propositional framework

Figure 6.12 Ranking of concepts from Sentence A

6.5.2 Summarization

This step is responsible for identifying the relevant propositions for the text. Here we use the VertexSort module of ExtroutNLP API, which adopts an empirical development model. This model applies quartiles associated with the graph topology to classify the vertices. This model is explained in Section 7.5.

From a directed graph with heavy vertices, the VertexSort model classifies each vertex in the *heavy*, *interjacent*, *adjacent*, and *light* classes. From observation, we define that the vertices associated with the first three types of classes are relevant for the text, these are defined as *relevant concepts*.

To identify the relevant propositions, we identify all propositions that are composed of relevant concepts. Therefore, a proposition (*concept1, relation, concept2*) is defined as *relevant proposition* if *concept1* and *concept2* are part of a relevant concepts list.

From of proposition list (Figure 6.10) and ranking of concepts (Figure 6.12), the following shows the relevant propositions identified.

(concept map, were developed in, course)
(course, is of, research program)
(research program, is of, american educator)
(american educator, sought to follow, change)
(american educator, sought to understand, change)
(change, include in, knowledge)
(knowledge, is of, child)
(researcher, interviewed, child)
(course, is of, study)
(researcher, found, course)
(study, to identify, change)
(change, include in, understanding)
(understanding, is of, child)
(understanding, is of, science concept)
(program, was based on, learning psychology)
(learning psychology, is of, american psychologist)
(idea, include in, psychology)
(psychology, is of, american psychologist)
(idea, is takes, place)
(place, is by, assimilation)
(assimilation, is of, concept)
(place, stay into, concept)
(science concept, is a, concept)
(research program, is a, program)
(learning psychology, is a, psychology)

Figure 6.13 Relevant propositions extracted from Sentence A

Following, in Figure 6.14, we present the concept map constructed for Sentence A (Figure 6.2) using the technological architecture proposed in the Section 6.2 .

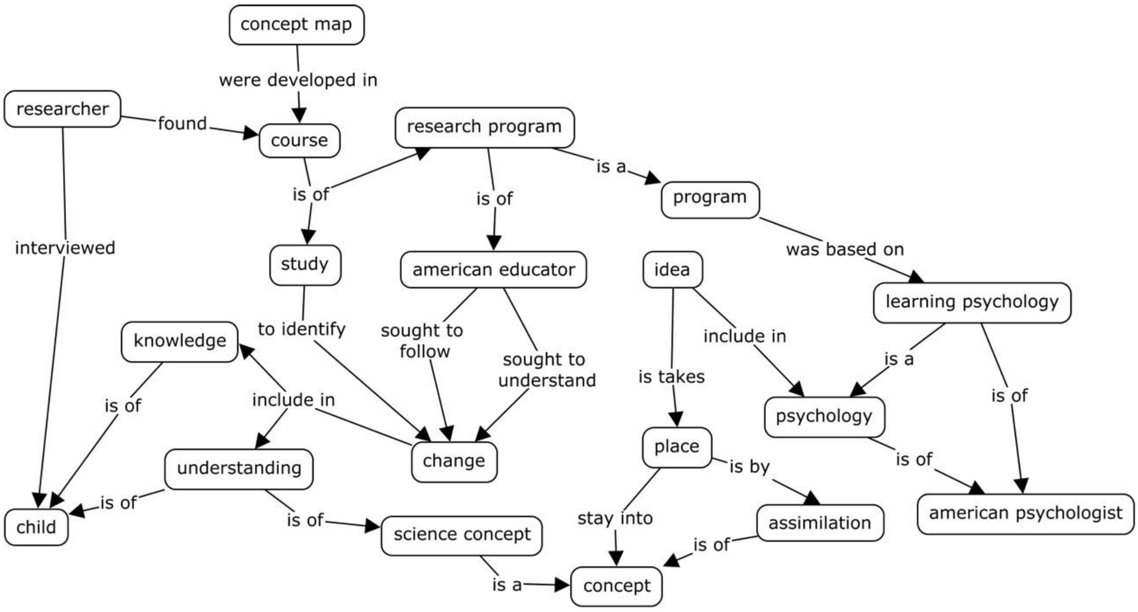


Figure 6.14 Concept map constructed from Sentence A

6.6 Service-Oriented Technological Architecture

In this section we present technological details of service-oriented architecture defined from the conceptual model proposed in Chapter 5. For the implementation of the service-oriented model, we adopt the following set of elements:

(i) **Protocol for data transmission:** The Hypertext Transfer Protocol (HTTP) is used to send the data of the requested service. The protocol is widely accepted and does not require additional access.

(ii) **Representation of data:** In order to enable a universal understanding of the messages exchanged between services and applications, we represent the data in

JavaScript Object Notation (JSON). JSON is in well-defined text format and completely independent of language.

(iii) **Publication of services:** The Representation State Transfer (REST), an architectural style based on HTTP protocol, is used to publish the services. REST uses URI to expose business logic and supports high-volume requests.

(iv) **Request of services:** since we use REST, then requests are performed for a specific URI using the methods provided by the HTTP protocol such as GET and POST.

6.7 Some Considerations on the Chapter

This chapter has proposed a technological architecture for the conceptual model presented in Chapter 5. This architecture is designed for a service-oriented model. Therefore, it was necessary to define components and resources adherent to this paradigm and to the premises defined in Section 6.1. This invalidated several components that were listed to compose the architecture. In addition, the choice of technical components was related to the quality of the concept map built. We chose to prioritize the components with greater accuracy over those with greater efficiency. Finally, the technical components for the Portuguese language required greater dedication of time, since we did not find components with satisfactory results and we dedicated some time to their improvement.

The next two chapters present productions developed from the technological architecture proposed in this chapter. Chapter 7 will present ExtroutNLP, a Java API for NLP tasks, and Chapter 8 introduces CMBuilder, a web tool for the automatic construction of concept maps from text.

Chapter 7
ExtroutNLP: Suite of Texts Processing Libraries

This chapter presents ExtroutNLP, a Java API that provides a suite of text processing libraries. This API is one of the research results and was motivated by the need found in the implementation of the conceptual model presented in Chapter 5.

This chapter is organized as follows: Section 7.1 presents the ExtroutNLP; Sections 7.2, 7.3, 7.4 and 7.5 describe the GenitiveInterpretation, the OpenIE, the Ranking, the VertexSort libraries; and Section 7.6 presents some considerations on the chapter.

7.1 About ExtroutNLP

ExtroutNLP is a Java API that provides a suite of text processing libraries, such as triple extraction, ranking and summarization for the Portuguese and English languages. This API uses common NLP tasks from the other toolkits, such as CoreNLP and OpenNLP, to provide more specific tasks for information extraction. Its goal is to make it easy to apply a number of information extraction libraries to a piece of text.

ExtroutNLP can be downloaded via the link

http://extroutnlp.lied.inf.ufes.br .

This link contains information, documentation, services and download of the API. The API will download a large zip file containing *(i)* the ExtroutNLP code jar, *(ii)* the models jar used by ExtroutNLP and *(iii)* the libraries required to run ExtroutNLP.

Alternatively, ExtroutNLP is available as a service and can be used through a default URI followed by the library and, subsequently, by the required service. All available services are listed in Section 5.2.2.1 and can be requested by a URI as using

http://extroutnlp.lied.inf.ufes.br/ExtroutNLP/<*lib*>/<*service*>.

ExtroutNLP, as well as all other libraries it uses, is licensed under the GNU General Public License v3. The license is free and allowed only for non-commercial uses.

The API is, initially, composed of four libraries, *GenitiveInterpretation*, *OpenIE*, *Ranking* and *VertexSort*, which will be presented in this chapter.

7.2 GenitiveInterpretation Library

The genitive interpretation is performed by an algorithm based on Dependence Graph tree. For this, we use the Dependency Parsing (CHEN & MANNING, 2014) module of

CoreNLP, a transition-based parser powered by a neural network which accepts word embedding inputs. The dependencies provide a representation of grammatical relations between words in a sentence in the format *DependencyType*(*governor*, *dependent*).

Figure 7.1 shows some dependencies extracted by Dependency Parse from a sentence. These dependencies can be represented by relations *nmod:poss*(program, Novak); *case*(Novak, ‘s); *nmod:poss*(knowledge, children); *case*(children, ‘s) and *nmod:of*(knowledge, science).

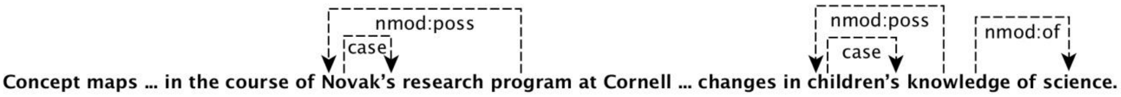

nmod:poss
case
nmod:poss
case
nmod:of
Concept maps ... in the course of Novak’s research program at Cornell ... changes in children’s knowledge of science.

Figure 7.1 Dependencies identified from a sentence

The algorithm searches for all *nmod:poss* dependencies on the tree. Then, it searches for all *case* dependencies formed by the dependent "*'s*" and for governor equal to dependent *nmod:poss* dependency identified. When this premise is satisfied, it removes the governor and dependent from the *case* dependency, i.e., “*Novak*” and “*’s*”, and adds the token "*of*" and governor of *case* dependency after governor of *nmod:poss* dependency.

7.3 OpenIE Library

The OpenIE is performed by an algorithm which implements a model for extracting open information based on linguistic structure. The model uses the linguistic structure of a text to identify triples in the format *subject-predicate-object*, compatible with propositions of the format (*concept1, relation, concept2*).

The model is based on constituent parse trees and consists of three steps: Independent Structures Identification, Structure Adjusting and Triples Extracting. These steps are explained in the course of this section.

7.3.1 Independent Structures Identification

During the first step of our method we apply segmentation on the parse tree in order to create a set of complete independent structures containing a less complex structure.

We define by complete independent structure, that formed by complete syntagms following Pattern 7.1 or 7.2 below. The complete syntagms are: *(i) NP syntagm*, contains a NN core or derivatives; *(ii) VP syntagm*, contains a VB core or derivative and a NP complete syntagm; and *(iii) PP syntagm*, contains an IN core or derivative and a NP complete syntagm. Intermediate structures, incomplete syntagmas and tags, existing among the complete syntagms of patterns are ignored.

$$S < ((NP < (NN+)) \$ (VP < (VB+ \$ (NP < (NN+))) \tag{7.1}$$

$$S < ((NP < (NN+)) \$ (PP < (IN \$ (NP < (NN+))) \tag{7.2}$$

This problem is treated as a depth-first search in the parse tree. The sequence in which vertices are visited is used to identify the patterns. The search starts top-down from the root node of the tree, and can take four actions:

(i) **Create structure:** an independent structure is created when the Patterns 7.1 or 7.2 are found.

(ii) **Recursive search top-down:** from a vertex v in *V*, it visits recursively all of its children vertices, until it finds one of the patterns or reaches a leaf node.

(iii) **Recursive search bottom-up:** from a vertex v in *V*, it visits recursively all of its ancestor vertices, until it finds one of the patterns or reaches the root node.

(iv) **Stop:** it stops when finding a leaf node in top-down or a root node in bottom-up search.

With these four actions, the search becomes a sequence of recursive actions top-down and bottom-up, followed by the action of creating a structure or to stopping. To perform the search, the model adopts four structure patterns, shown in Table 7.1:

Pattern	*Structure*	*Sentence Example*
I	(S (NP) (VP (NP)))	**(S (NP** (NN concept) (NN map)**)(VP** (VBP are) **(NP** (NP (JJ graphical) (NN tool)))**)** [...]
II	(NP) (S (VP (NP)))	**(NP** (NP (NN word)) (PP (IN on) (NP (DT the) (NN line))) (, ,) (VP (VBN referred) (PP (TO to)) (PP (IN as) **(S(VP** (VP (VBG linking) **(NP** (NN word)**))** (CC or) **(VP** (VBG linking) **(NP** (NN phrase)**))))**) [...]
III	(NP (NP) (VP (NP)))	**(NP (NP** (DT a) (VBG connecting) (NN line)) **(VP** (VBG linking) **(NP** (CD two) (NN concept)**)))** [...]
IV	(NP) (PP (NP))	**(NP** (NN concept)) (, ,) (VP (ADVP (RB usually)) (VBN enclosed) **(PP** (IN in) **(NP** (NP (NN circle) (CC or) (NN box))**))** [...]

Table 7.1 Structure Patterns

These patterns allow for the identification of a complete structure beyond a simple declarative syntagm clause (S), enabling it to inherit the subject or complement of other clauses.

From these four explained actions and from patterns defined in Table 7.1, the search extracts the complete independent structures adopting the following rules:

(i) **Pattern I:** In the S syntagm found, one needs to check if there is a complete NP and VP syntagm. If there is a sequence of incomplete NPs or VPs, the search will go deep into the tree to find complete syntagm (Recursive search top-down). If found, the structure is extracted (Create structure).

(ii) **Pattern II:** If the NP syntagm is missing or incomplete and the VP is complete, one needs to search for a complete NP syntagm in ancestors (Recursive search bottom-up). If an ancestor NP syntagm is found, the structure is extracted (Create structure).

(iii) **Pattern III:** In the NP syntagm found, one needs to check if there is a complete NP and VP syntagm. If there is a sequence of incomplete NPs or VPs, the search will go deep into the tree to find complete syntagm (Recursive search top-down). If found, the structure is extracted (Create structure).

(iv) **Pattern IV:** In the PP syntagm found, one needs to check whether the PP syntagm is complete. If so, one needs to search for a complete NP syntagm in ancestors (Recursive search bottom-up). If an ancestor NP syntagm is found, the structure is extracted (Create structure).

7.3.2 Structure Adjusting

This step adjusts the integral parts of an independent structure, without harming the context. The following shows the tasks applied to adjust the structures:

Identify Syntagm Nucleus: The identified nucleus will be used as label for concepts and relationships. To identify the syntagm nucleus that makes up the structure, we use the following patterns defined in Table 7.2. Only the tags defined in Table 3.1 (Chapter 3) are used, all others are removed.

Syntagm	***Pattern***	***Example***
NP	[NN+]	(NN concept) (NN map)
NP	[JJ?] [NN+]	(JJ graphical) (NN tool)
NP	[VB+] [NN+]	(VBG connecting) (NN line)
VP	[VB+]	(VBN connected) (VBG using)
VP	[VB+] [RB?]	(VBP are) (RB not)
VP	[VB+] [IN?]	(VBN enclosed) (IN in)
PP	[IN+]	(IN between)

Table 7.2 Patterns to identify syntagm nucleus

Lemmatize: All tokens that belong to syntagm nucleus of a Nominal Phrase are lemmatized. For this, we use module Lemmatization of CoreNLP. Lemmatization was not performed in previous steps because it could alter the syntactic function of the words in the sentence.

Convert Prepositional Syntagm: Every prepositional syntagm is transformed into a verbal syntagm. For this, we created a VP syntagm formed by content of the PP syntagm and a verbal token added before the first prepositional token. This verbal token is assigned according to Table 7.3:

Preposition	***Token***
[IN of] | [IN for] | [IN to] | [IN by] | [IN from]	[VB is]
[IN in]	[VB include]
[IN on] | [IN at] | [IN into]	[VB stay]
[IN as] | [IN between]	[VB appear]
[IN about]	[VB refer]
[IN with]	[VB have]
[IN de] | [IN para] | [IN pelo] | [IN desde] | [IN por] | [IN a partir de]	[VB é]
[IN em]	[VB incluir]
[IN dentro]	[VB permanecer]
[IN como]	[VB aparece]
[IN sobre]	[VB refere]
[IN com]	[VB ter]

Table 7.3 Verbal tokens mapped to prepositions

Noun Similarity: All NN tokens of NP syntagm have their similarity analyzed in order to assign a unique noun to similar ones. To calculate the similarity of each NN token with all other NN tokens existing, we use the similarity measure LIN and multi-word structure. This measure is calculated by the API WS4J and based on WordNet provided by API JawJaw. If the similarity between two NN tokens is greater than 0.8, then the NN token appearing less frequently in the text is replaced by the one with the highest frequency.

Identify Specialization Relationships: a specialization relationship is identified from of compound name or grammatical structure. When a nucleus A, formed by more than one noun, contains a nucleus B, formed by only one noun of nucleus A, a specialization relationship is identified between nouns.

The following patterns (Table 7.4) are used to identify the relations between nouns:

Pattern	***Remaking***	***Example***	***Triple***
(**JJ**) (**NN**)	(**NN**) has property (**JJ**)	(JJ graphical) (NN tool)	(toll, has property, graphical)
(**NN0**+**NN1**)	(**NN0**+**NN1**) is a (**NN1**)	(NN blackbird)	(blackbird, is a, bird)
(**NN0**) (**NN1**)	(**NN0**)(**NN1**) is a (**NN1**)	(NN research) (NN program)	(research program, is a, program)
(**NN**) (**NNP**)	(**NNP**) is a (**NN**)	(NN psychologist) (NNP david) (NNP ausubel)	(david ausubel, is a, psychologist)

Table 7.4 Patterns to identify specialization relationships

In addition to these patterns, other specialization relations are extracted with the triple extract step from the independent structures. The following structures (Table 7.5) are used to identify the relations in the structures:

Structure	***Remaking***	***Example***	***Triple***
(S (NP (**NN0**)) (VP (VB) (NP (**NN1**) (CC) (**NN2**))	(**NN0**) (VB) (**NN1**) (**NN0**) (VB) (**NN2**)	(S (NP (NNS record)) (VP (VBZ likes) (NP (NNS event) (CC or) (NNS object))))	(record, likes, event) (record, likes, object)
(S (NP (**NN0**)) (PP (JJ) (IN) (NP (**NN1**) (,) (**NN2**) (CC) (**NN3**))))	(**NN0**) (VB) (IN) (**NN1**) (**NN0**) (VB) (IN) (**NN2**) (**NN0**) (VB) (IN) (**NN3**)	(S (NP (NNS symbol)) (PP (JJ such) (IN as) (NP (NN person) (CC or) (NN image))))	(symbol, appear as, person) (symbol, appear as, image)

Table 7.5 Structures to identify specialization relationships

Convert Named Entities: The named entities (places, organizations and proper names) are not directly represented, since we understand that they are instances of classes (concepts) and should not be represented on the map. However, identifying their types can be of interest to our concept map.

To this end, we use the Named Entity Recognition (FINKEL, et al., 2005) module of CoreNLP, a general implementation of linear chain Conditional Random Field (CRF) sequence models. The Named Entity Recognition is applied to the existing text in the structures. Each structure containing the named entity is retained in a textual summary.

With the support of Jena API, a query containing the entity label and the entity type is executed on DBPedia. The description of named entity in the DBPedia, and the textual summary is stored as a vector representation. The similarity between the vectors is calculated with the Cosine similarity, supported by API DKPro. If similarity is greater than 0.8, the named entity token is replaced by a token containing a named entity description in DBPedia. Otherwise, it is replaced by a token containing named entity type.

An example of a SPARQL query supported by Jena API is shown below, where the variable "*var*" is replaced by a named entity, for example "*Novak*" and return its description, "*American educator*".

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

Query SPARQL

7.3.3 Extract Triples

This step aims to extract triples to represent the facts expressed in complete independent structures, i.e., the propositions. The triples must be meaningful, represent the faithfully the information (explicit or implicit) and express as many facts as possible in their smallest meaning unit.

To identify the constituent parts of an independent structure and consequently extract its triples, we adopted four general rules:

(i) We located the first VP syntagm in the independent structure. From it, we define subject as the NP syntagm located before of VP syntagm; object as the NP syntagm located within VP syntagm; and predicate as the structures located between the VP syntagm and the object.

(ii) In the subject, we seek for other NP syntagms. If they exist, the nucleus of each NP will result in a distinct subject. Otherwise, the syntagm nucleus will result in a single subject.

(iii) In the predicate, we seek for other VP syntagms. If they exist, each syntagm will compose a part of common predicate, until the penultimate VP identified. In the last VP, we search for all the syntagm nuclei. Each VP nucleus will be associated with the common predicate and will result in a distinct predicate.

(iv) In the object, we seek for other NP syntagms. If they exist, the nucleus of each NP will result in a distinct object. Otherwise, the syntagm nucleus will result in a single object.

These rules will produce a set of subjects, predicates, and objects for each structure. From them, we combine all possibilities to create a set of triples in the form (*argument1, relation, argument2*) that represents a proposition in format (*concept1, relation, concept2*).

Finally, all triples extracted from the text are reviewed. If there are repeated triples, only one of them is kept and all others are excluded.

7.3.4 Experiments using OpenIE

In order to check the quality of propositions extracted from OpenIE module of ExtroutNLP library, we compare it against the two Open IE systems: OLLIE and ClausIE (Section **3.2.3.5**). These systems are based on dependency parser, unlike ExtroutNLP which is based on constituent parser. Other systems based on the constituent parser were not included because they are ancestors of OLLIE, such as ReVerb.

7.3.4.1 Example Extractions

Following we illustrate the differences between the extractors for a manually-selected example sentence. Table 7.6 shows our evaluation and the propositions extracted by each

Open IE extractor for the sentence: "*He fathered two children, Edna and Donald, and lived in Aberdeen until his death from tuberculosis in 1942.*".

Extractor	***Id***	***Proposition***	***Evaluation***
OLLIE	O1	("He"; "lived in"; "Aberdeen")	Correct
	O2	("He"; "lived until"; "his death")	Correct
	O3	("He"; "fathered"; "two children")	Correct
ClausIE	C1	("He", "fathered", "two children")	Correct
	C2	("He", "fathered", "Edna")	Correct
	C3	("He", "fathered", "Donald")	Correct
	C4	("He", "lived", "in Aberdeen")	Correct
	C5	("He", "lived", "in Aberdeen until his death")	Correct
	C6	("He", "lived", "in Aberdeen from tuberculosis in 1942")	Incorrect
	C7	("his", "has", "death")	Correct
ExtroutNLP	E1	("He", "fathered", "two child")	Correct
	E2	("He", "fathered", "Edna")	Correct
	E3	("He", "fathered", "Donald")	Correct
	E4	("He", "lived in", "Aberdeen")	Correct
	E5	("He", "lived until", "death")	Correct
	E6	("death", "is from", "tuberculosis")	Correct

Table 7.6 Example extractions for each extractor

In the sequence, we highlight relevant points presented by the extracted propositions in the table:

(i) OLLIE extracts few propositions from the text and does not cover the whole context of the sentence.

(ii) ClausIE extracts some duplicate propositions as C5 and C6, it extracts extensive arguments such as whole predicate, and uses possessive pronoun as argument (C7).

(iii) ExtroutNLP does not identifies some information such as "in 1942", it loses the context of some propositions by dividing them (E6).

We also highlight two main characteristics that differentiate ExtroutNLP: it does not extract duplicate propositions, i.e., several propositions from the same core; and it divides the predicate into smaller propositional units.

7.3.4.2 Analysis and Results

The extractors were evaluated on one dataset commonly used for this purpose containing 60 random sentences from the English Wikipedia, available on the web site[4].

The labels of their evaluation are also available by the dataset. However, we chose to label all extracted triples (by all three extractors) for the following reasons: there are different

[4] www.mpi-inf.mpg.de/departments/d5/software/clausie

interpretations, to avoid errors by labeling triple at different times, and to analyze the characteristics of the ExtroutNLP.

Therefore, each sentence was processed by the three extractors generating a list of propositions. We labeled each proposition with a label for *accuracy* (yes or no). This evaluation is available on the web site[5].

Table 7.7 shows the results calculated to the dataset comparing the extractors. Moreover, we computed the triple length, i.e., its average number of words.

Information	*OLLIE*	*ClausIE*	*ExtroutNLP*
All Propositions Extracted	183	326	339
Correct Propositions Extracted	107	188	221
Precision	0.58	0.57	0.65
Triple Length	9.39	10.80	5.11

Table 7.7 Results of the evaluation for the extractors

Analyzing the table, we noticed that all extractors have large number of propositions extracted and correct. However, we observe that:

(i) OLLIE extracts few propositions, 1/3 less than the others. The accuracy is high, even though the number of propositions extracted is low, i.e., it has lower coverage.

(ii) ClausIE and ExtroutNLP extract similar number of propositions. However, ClausIE, as discussed earlier, extracts many duplicate propositions. Such propositions received *yes* label for accuracy in our evaluation.

(iii) ClausIE extracts propositions containing the largest number of words, unlike ExtroutNLP that extracts the smallest propositional units.

(iv) OLLIE and ClausIE does not solve anaphora, unlike ExtroutNLP.

(v) Although OLLIE (MaltParser) and ClausIE (Stanford) use dependency parser and ExtroutNLP (Stanford) use consistency parser, both extractors obtained good results.

We can consider that the results presented by the extractors were satisfactory, since the dataset is composed by sentences with characteristics difficult to treat, such as short sentences, possessive cases, passive voice, explanatory sentence, apposed, entity, anaphora among others.

From the experiments, we identified some critical points related to ExtroutNLP. These points can be observed from the sentence "*Daughter of the actor Ismael Sanchez Abellan and actress*

[5] extroutnlp.lied.inf.ufes.br/resources

and writer Ana Maria Bueno (better known as Ana Rosetti), Gabriel was born in San Fernando, Cadiz, but spent her childhood in Madrid.":

(i) **Apposed**: it does not identify. The text "Daughter of the actor Ismael Sanchez Abellan and actress and writer Ana Maria Bueno (better known as Ana Rosetti)" is not associated to Gabriel.

(ii) **Anaphora**: it solves some anaphora wrongly. The word "*her*" was associated with "*Ana Maria Bueno*".

(iii) **Incomplete Proposition**: some propositions were splitted up into smaller units and lost its context, such as "*childhood in Madrid*" into (*Daughter, spent, childhood*) and (*childhood, in, Madrid)*

7.4 Ranking Library

This library implements an empirically developed measure for ranking concepts, i.e., it assigns a weight to concepts and then orders them. The measure developed is called HAF, based on concepts of hub and authority vertex (KLEINBERG, 1999) associated with the frequency of concept in the text and in the domain as explained in the following sections.

7.4.1 HAF Model

The model treats the propositions as a graph, where the vertex represents a concept and each edge represents relation between concepts. From the propositions presented in Figure 6.10, a graph as shown in Figure 7.2 is built.

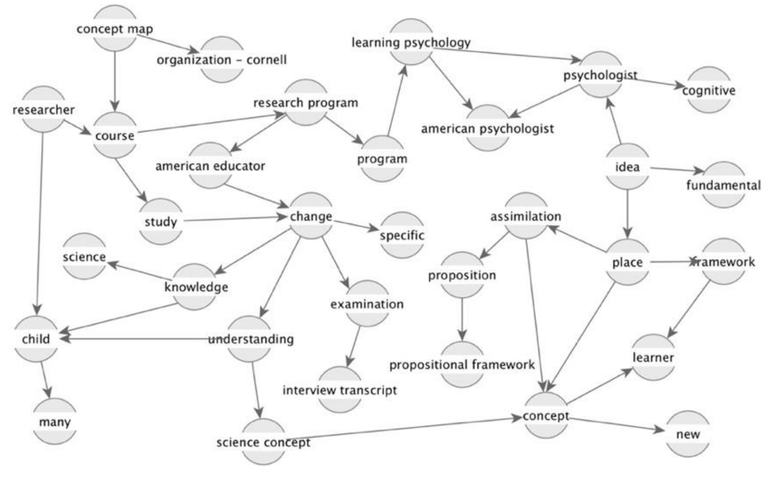


Figure 7.2 Representation of propositions in a graph

Following the theory presented by (KLEINBERG, 1999), each vertex of the graph has a *hub* and *authority* score. The authority vertex contains valuable information on the subject and hub vertex contains useful links towards the authoritative vertex (SHATAKIRTI, 2011), as exemplified in Figure 7.3.

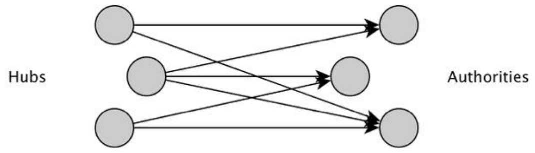


Figure 7.3 Representation of hub and authorities vertex

The HAF model associates the authority and hub concepts with the frequency of concepts in the text, based on the Hub-Authority-Root-Distance (HARD) model. The HAF model estimates the concept importance based on the following factors: *(i)* authority value, number of incoming connections; *(ii)* hub value, number of output connections; *(iii)* frequency of the concept in the document; and *(iv)* frequency of the concept in the domain.

The weight *W* of each concept *k*, is computed by the following formula:

$$\boldsymbol{W(k) = [\beta.TFIDF_{\Omega}(k)] + [\alpha.TF_{d}(k)] + [\gamma.(\rho.A(k) + \sigma.H(k))]} \qquad (7.3)$$

In the formula, *TF-IDF*$_{\Omega}$ is the inverse frequency of the concept in the domain (Section 3.1.2.1), *TF*$_{d}$ is the frequency of the concept in the document (Section 3.1.2.1), *A* is the weight of the authoritative nodes, and *H* is the weight of the hub nodes.

The best parameters adjustment to authority and upper nodes in the HARD model (REICHHERZER & LEAKE, 2006) were assigned to $\rho = 2.235$ and $\sigma = 1.764$ in the HAF model. The parameters $\beta = 0.1$, $\alpha = 0.2$, and $\gamma = 0.7$ were adopted in the experiment step. The maximum weight calculated by HAF model for a vertex is equal to 1.

7.4.2 Experiments using Ranking

We consider that the graph topology can direct the identification of relevant concepts in a concept map. Based on this hypothesis, we started the experiments using graph-based algorithms and models to identify which best represents the ranking as we want.

For our purpose, we want a ranking composed of the most prominent concepts in the graph topology, i.e., those that are central to the construction and connection of a concept map. Following are results of some algorithms and models studied.

HITS Algorithm: described in Section 3.1.2.3, was implemented using the JUNG library. On the graph represented in Figure 7.2, the algorithm is applied in its classic behavior taking interactions = 100 and tolerance = 0.00001. Figure 7.4 shows the ranking of concepts constructed by the algorithm, where weight is the sum of the authority plus hub value.

Ranking	Weight	Concepts	Ranking	Weight	Concepts
1	0.5904	place	17	0.1815	examination
2	0.5828	concept	18	0.1772	science
3	0.5808	assimilation	19	0.1708	specific
4	0.5191	child	20	0.1543	american psychologist
5	0.4813	understanding	21	0.1179	fundamental
6	0.4813	knowledge	22	0.1162	organization - cornell
7	0.4234	science concept	23	0.0761	cognitive
8	0.3906	researcher	24	0.0747	research program
9	0.3525	change	25	0.0718	learner
10	0.3248	course	26	0.0651	study
11	0.2921	psychology	27	0.0651	american educator
12	0.2813	framework	28	0.0555	program
13	0.2093	idea	29	0.0526	new
14	0.2058	concept map	30	0.0359	many
15	0.1958	proposition	31	0.0359	propositional framework
16	0.1883	learning psychology	32	0.0359	interview transcript

Figure 7.4 Ranking of concepts constructed from HITS algorithm

PageRank Algorithm: described in Section 3.1.2.3, was implemented using the JUNG library. On the graph represented in Figure 7.2, the algorithm is applied taking interactions = 100 and tolerance = 0.00001. The Figure 7.5 shows the ranking of concepts constructed by the algorithm.

Ranking	Weight	Concepts	Ranking	Weight	Concepts
1	0.0635	change	17	0.0270	science concept
2	0.0604	many	18	0.0270	science
3	0.0583	learner	19	0.0254	study
4	0.0545	concept	20	0.0254	program
5	0.0498	american psychologist	21	0.0254	research program
6	0.0477	child	22	0.0254	course
7	0.0413	interview transcript	23	0.0254	american educator
8	0.0400	new	24	0.0219	proposition
9	0.0381	learning psychology	25	0.0191	organization - cornell
10	0.0360	psychology	26	0.0184	assimilation
11	0.0346	propositional framework	27	0.0184	framework
12	0.0307	cognitive	28	0.0169	place
13	0.0286	specific	29	0.0169	fundamental
14	0.0286	examination	30	0.0127	idea
15	0.0286	understanding	31	0.0127	concept map
16	0.0286	knowledge	32	0.0127	researcher

Figure 7.5 Ranking of concepts constructed from PageRank algorithm

HARD Model: associates weights to nodes based on its authority value, hub value, and upper node value (shortest distance to root concept) (LEAKE, et al., 2004). The weight *W(k)* of each concept *k* is computed by the Formula 7.4, where *A(k)*, *H(k)*, and *U(k)* are the authority, hub and upper node values for *k*. The parameters assigned for $\rho = 0$, $\sigma = 2.235$ and $\phi = 1.764$ were found in the best adjustment made by (LEAKE, et al., 2004).

$$\boldsymbol{W(k) = [\rho . A(k) + \sigma . H(k) + \phi . U(k)]} \qquad (7.4)$$

The HARD model is applied to the graph represented in Figure 7.2. The Figure 7.6 shows the ranking of concepts, where the upper node value and parameter p is 0.0.

Ranking	Weight	Concepts	Ranking	Weight	Concepts
1	8.9400	change	17	2.2350	examination
2	6.7050	idea	18	2.2350	proposition
3	6.7050	place	19	2.2350	framework
4	4.4700	concept	20	2.2350	american educator
5	4.4700	concept map	21	2.2350	child
6	4.4700	psychology	22	0.0000	many
7	4.4700	research program	23	0.0000	american psychologist
8	4.4700	course	24	0.0000	specific
9	4.4700	understanding	25	0.0000	organization - cornell
10	4.4700	knowledge	26	0.0000	cognitive
11	4.4700	researcher	27	0.0000	propositional framework
12	4.4700	assimilation	28	0.0000	interview transcript
13	4.4700	learning psychology	29	0.0000	new
14	2.2350	study	30	0.0000	fundamental
15	2.2350	science concept	31	0.0000	learner
16	2.2350	program	32	0.0000	science

Figure 7.6 Ranking of concepts constructed from HARD model

HAF Model: described in Section 7.4.1, was implemented using the ExtroutNLP library. In this experiment we do not use the frequency of concepts in the domain, since this parameter is variable according to the knowledge base. Therefore, on the graph shown in Figure 7.2, we apply the reduced HAF model:

$$\boldsymbol{W(k) = \ [\alpha . TF_d(k)] + [\gamma . (\rho . A(k) + \sigma . H(k))]} \qquad (7.5)$$

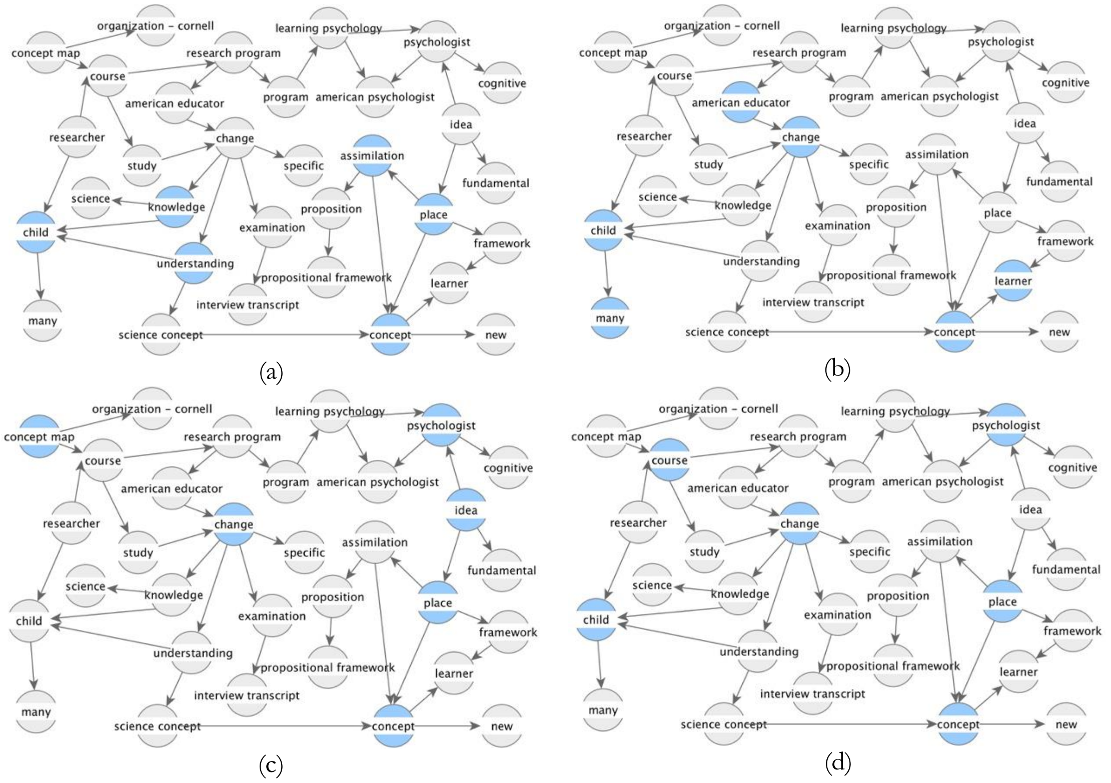
Figure 7.7 shows the ranking of concepts constructed by the model.

Ranking	Weight	Concepts	Ranking	Weight	Concepts
1	0.8333	change	17	0.3176	concept map
2	0.6814	concept	18	0.2912	science concept
3	0.6490	course	19	0.2912	program
4	0.6226	child	20	0.2912	examination
5	0.5421	place	21	0.2912	proposition
6	0.5157	psychology	22	0.2647	learner
7	0.4431	idea	23	0.2245	framework
8	0.4167	research program	24	0.1657	fundamental
9	0.4167	understanding	25	0.1657	organization - cornell
10	0.4167	knowledge	26	0.1657	new
11	0.4167	assimilation	27	0.1657	many
12	0.4167	learning psychology	28	0.1657	specific
13	0.3843	researcher	29	0.1657	interview transcript
14	0.3579	study	30	0.1657	cognitive
15	0.3579	american educator	31	0.1657	science
16	0.3314	american psychologist	32	0.0990	propositional framework

Figure 7.7 Ranking of concepts constructed from HAF model

7.4.2.1 Analysis of the Experiments

To better analyze the ranking constructed by the algorithms and models, we direct our observations to the top 6 concepts of each ranking, that is, 20%. The Figure 7.8 highlights the top 6 concepts of each ranking in the graph: HITS (a), PageRank (b), HARD (c) and HAF (d).



(a) (b) (c) (d)

Figure 7.8 Top concepts of the ranking

Following are some characteristics observed on the graphs:

(i) The HITS algorithm prioritizes the relevant concepts in the topology of the graph and its neighbors. This case can be observed in the *place* and *child* concepts, they have relevance in the graph and influence the weight of the neighboring concepts such as *concept*, *assimilation*, *understanding* and *knowledge*.

(ii) The PageRank algorithm, although it contains different feature of HITS, also prioritizes concepts that are neighbors to the relevant concepts in the graph topology. This case can be observed in the *change, concept* and *child* concepts, they have relevance in the graph and influence the weight of the neighboring concepts such as *American educator*, *many* and *learner*. These neighbor concepts have little relevance in the topology of the graph.

(iii) The HARD model prioritizes relevant concepts in the graph topology based only on output connections of the vertices.

(iv) The HAF model prioritizes relevant concepts in the graph topology based on frequency of the concepts and incoming/output connections of the vertices.

Looking at the experiments, we conclude that HITS or PageRank strategies can create more concise maps, however can be less broad. In contrast, HARD or HAF model strategies can create broader maps, however can be less concise.

In order to create a map that comprehensively represents the subject of the text, we adopt the HARD and HAF models in our studies. For this purpose, we consider the HAF model more appropriate, since it has selected the most central vertices of the graph in a balanced way.

7.5 VertexSort Library

This library implements an empirically developed model for classifying the vertices of a directed graph. The weight of each concept, calculated by some ranking, is attributed to its corresponding vertex in the graph. The following (Figure 7.9) shows the graph containing the weight of the vertices according to the ranking calculated by the HAF model in Figure 7.7.

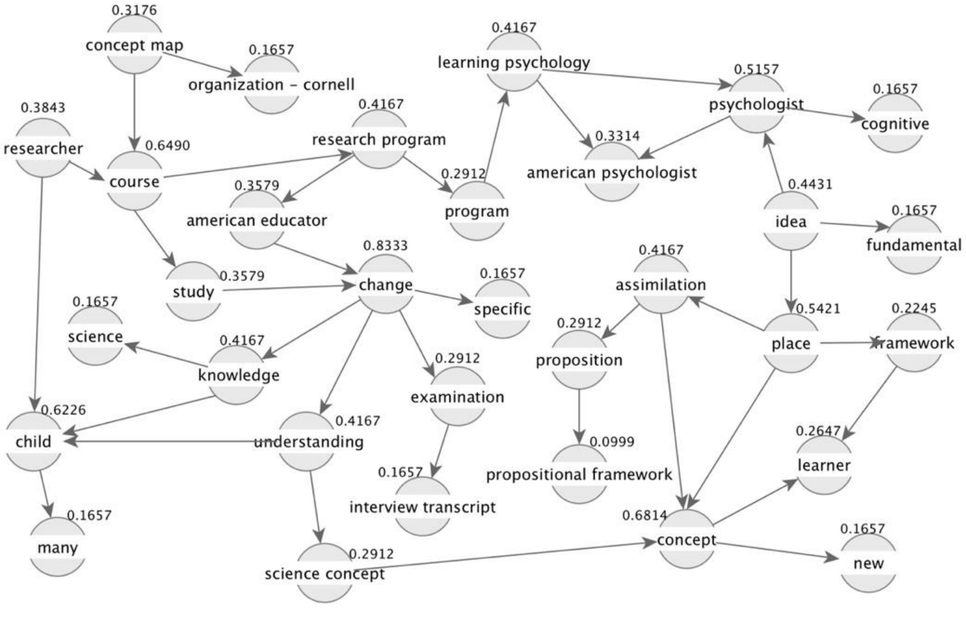


Figure 7.9 Graph containing the weight of the vertices

The model developed defines four classes to a vertex: *Heavy*, *Interjacent*, *Adjacent* and *Light*, as explained in the following.

(i) **Heavy vertex:** is defined as the most relevant vertex in the graph. These vertices are identified by applying quartiles in order to divide the distribution of the concepts ranking into four equal parts. For the identification of heavy vertex we

adopt the third quartile or upper quartile in order to identify 25% highest weights in the ranking.

(ii) **Interjacente Vertex:** is identified by all intermediate nodes that are in the path between two heavy vertices, that is, if a heavy vertex has an output connection to a N vertex and this vertex has an output connection to a heavy vertex or to other vertices who have an output connection to a heavy vertex, all vertices in that path are classified as interjacent.

(iii) **Adjacent Vertex:** is defined as vertices that are in the vicinity of heavy and interjacent vertex and having a weight exceeding a threshold. This limit is defined as the lowest weight presented by interjacent vertices. The neighboring vertices are defined as *(i) input*, a vertex sending a connection to a heavy or interjacent vertex and all intermediate vertices up to the first vertex of the path; and *(ii) output*, a vertex that receives a connection coming from a heavy or interjacent vertex and all intermediate vertices up to the last vertex of the path. The weight for the input and output vertex is the average sum of all intermediate vertices of its path. Thus, the input and output neighbors vertex containing weight greater or equal to the threshold is classified as adjacent, as well as all intermediate vertices of its path.

(iv) **Light Vertex:** is defined as the least relevant vertex in the graph. All other vertices not classified as heavy, interjacent or adjacent are classified as light.

On the graph represented in Figure 7.9, we highlight in Figure 7.10 the Heavy, Interjacent, Adjacent and Light vertices identified by the model in blue, pink, green and yellow colors, respectively.

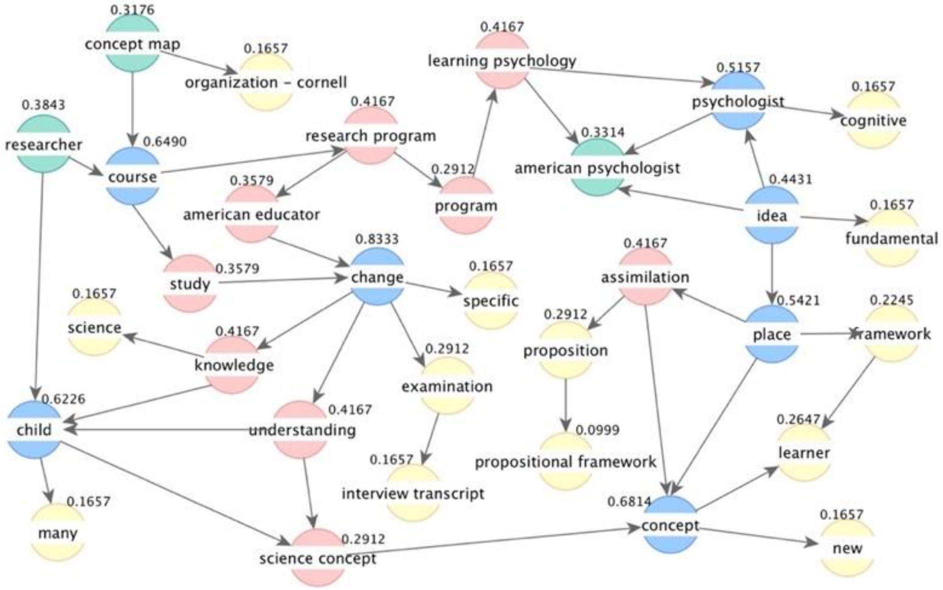


Figure 7.10 Graph representing the vertices class

Analyzing the classification of the vertices presented in Figure 7.10, we can observe: *(i)* the interjacent vertices are located between the first and third quartiles containing average weight in the ranking of concepts (Figure 7.7); *(ii)* on the graph presented was defined 0.2912 as threshold to adjacent vertex; *(iii)* although the proposition and examination vertices contain weight greater than the adjacent threshold, they were classified as light vertex, because the weight of their path is less than the adjacent threshold.

7.6 Some Considerations on the Chapter

This chapter presented ExtroutNLP, a Java API composed, initially, of four libraries directed to information extraction tasks in texts. This API was developed from the need of technological components to perform tasks of NLP and Information Extraction required in the conceptual architecture, presented in Chapter 5.

OpenIE was necessary because our goal is to extract appropriate propositions for building maps, that is, proposition formed by smaller units and without duplicity. Ranking and VertexSort are attempts to create more appropriate methods for building maps.

Besides to developing ExtroutNLP, we provide components to be downloaded and embedded because our goal is to share the information and ensure that other projects can use, expand, or improve them.

Although some models have been prepared in an empirical way, in general, they were based on some study already realized in other domains. Thus, these models tend to be further studied for validation.

In the next chapter, we introduce CMBuilder, a tool developed from technological architecture using the API ExtroutNLP.

Chapter 8

CMBuilder: A Web Tool for the Automatic Construction of Concept Maps from Texts

This chapter presents the CMBuilder, a web tool for the automatic construction of concept maps from texts in Portuguese and English languages. This tool was built using the technological architecture presented in Chapter 6 and it provides a proof of the concept of the issues addressed by this research.

This chapter is organized as follows: Section 8.1 presents the CMBuilder; Section 8.2 describes how CMBuilder works; Sections 8.3 and 8.4 present and discuss some experiments with English and Portuguese language; and Section 8.5 shows some considerations on the chapter.

8.1 About CMBuilder

CMBuilder, the acronym for Concept Map Builder, is a web tool whose purpose is to automatically construct a concept map of scientific style from an academic text in Portuguese or English language.

The CMBuilder will be provided along with a service-based platform, CMPaaS (CURY, et al., 2014), presently under development in our laboratory. This platform aims at expanding and integrating basic services, such as edition, management, and manipulation of concept maps.

To date, this platform offers services for merging concept maps (VASSOLER, et al., 2014), information retrieval on maps from questions (PERIN, et al., 2014), and shallow ontologies construction from maps (PINOTTE, et al., 2015). In this context, the present research proposes the new service on the CMPaaS to construction of concept maps from texts (AGUIAR, et al., 2016).

Besides, the platform needs to be utilized with an application that provides a user interface, in this case, the Knowledge Portal. This portal is a final interface for the use of tools developed for the services provided by CMPaaS. The Figure 8.1 presents a conceptual architecture to illustrate where the CMBuilder tool and its components are inserted in that context.

CMBuilder is an implementation of technological architecture presented in Chapter 6, i.e., a tool with service-oriented approach. For this, we define the following specifications:

(i) For the Data Layer the Neo4J 3.0 graph database is adopted.

(ii) For the Services Layer we adopt the Java language with JSE 8 and Spring Framework 4.3.4.

(iii) For the Presentation Layer we adopt the Python language 2.7.9 and Django 1.6.5.

(iv) For services publication the Linux Ubuntu 14.04 server and the Tomcat 8.0.30 java web server are used.

(v) For implementation we use the object-oriented programming paradigm.

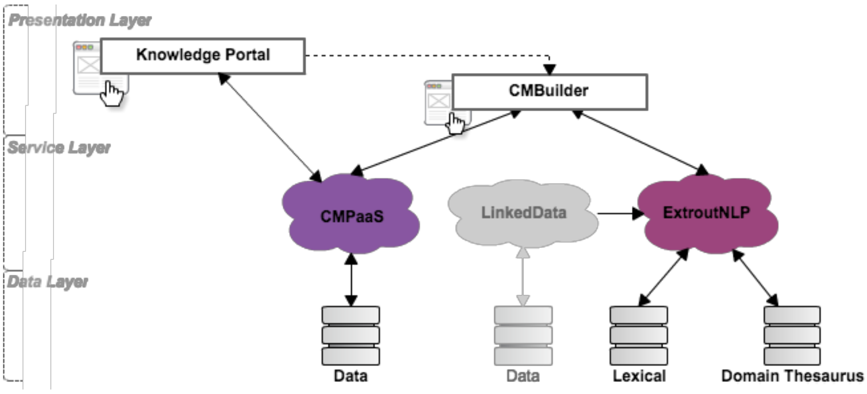


Figure 8.1 Conceptual architecture of context

This way, the CMBuilder tool is available for use through the Knowledge Portal and the ExtroutNLP API is available for expansion and embedding through a service. Thus, the CMBuilder can be accessed via the link

http://cmpaas.inf.ufes.br/cmbuilder .

The development of this tool includes three modules which have already been presented in this research: *Elements Extractor*, *Domain Identifier* and *Summarizer* which will be explained in the course of this chapter. The *Formatter* module has not been implemented so far, since different pdf formats and configurations have damaged the extraction and cleaning of the text.

8.2 CMBuilder Operation

The subsequent sections describe the operation of CMBuilder, i.e., the steps from the input of the data source to the construction of the concept map.

8.2.1 To Access

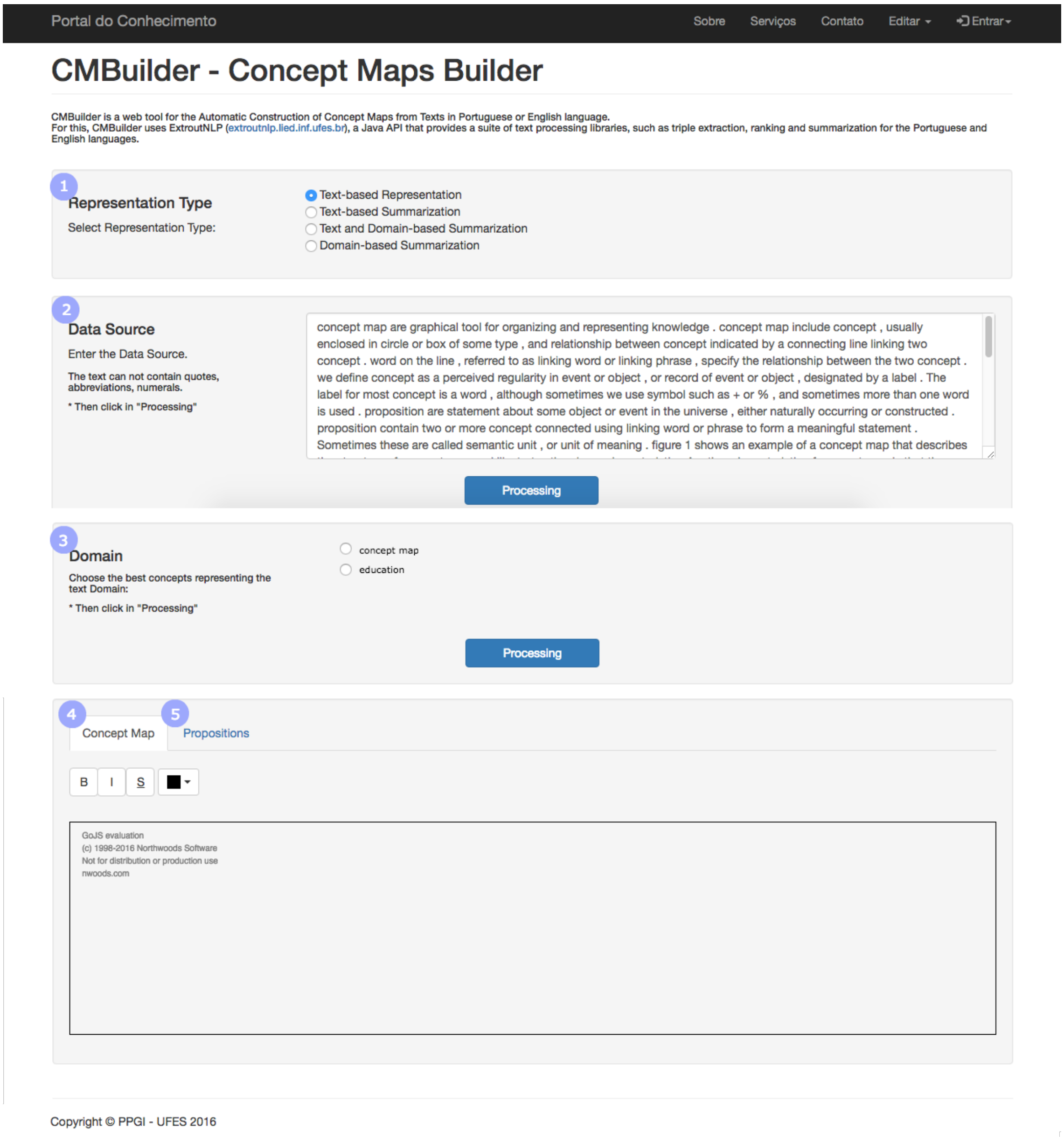
The user can access CMBuilder in the following two ways:

(i) Access through the Knowledge Portal by *http://cmpaas.inf.ufes* link and choose the CMBuilder service.

(ii) Access through the CMBuilder directly by the *http://cmpaas.inf.ufes.br/cmbuilder* link. The CMBuilder does not require login to the Knowledge Portal.

8.2.2 The Main Interface

The CMBuilder interface is shown in Figure 8.2. The interface consists of the following areas: Representation Type (1), Data Source (2), Domain (3), Concept Map (4) and Propositions (5) and the following user actions: Select Representation Type, Insert Text, Select Domain, and Process Text.



CMBuilder - Concept Maps Builder

CMBuilder is a web tool for the Automatic Construction of Concept Maps from Texts in Portuguese or English language.
For this, CMBuilder uses ExtroutNLP (extroutnlp.lied.inf.ufes.br), a Java API that provides a suite of text processing libraries, such as triple extraction, ranking and summarization for the Portuguese and English languages.

Representation Type

Select Representation Type:

● Text-based Representation
○ Text-based Summarization
○ Text and Domain-based Summarization
○ Domain-based Summarization

Data Source

Enter the Data Source.

The text can not contain quotes, abbreviations, numerals.

* Then click in "Processing"

concept map are graphical tool for organizing and representing knowledge . concept map include concept , usually enclosed in circle or box of some type , and relationship between concept indicated by a connecting line linking two concept . word on the line , referred to as linking word or linking phrase , specify the relationship between the two concept . we define concept as a perceived regularity in event or object , or record of event or object , designated by a label . The label for most concept is a word , although sometimes we use symbol such as + or % , and sometimes more than one word is used . proposition are statement about some object or event in the universe , either naturally occurring or constructed . proposition contain two or more concept connected using linking word or phrase to form a meaningful statement . Sometimes these are called semantic unit , or unit of meaning . figure 1 shows an example of a concept map that describes

Domain

Choose the best concepts representing the text Domain:

* Then click in "Processing"

○ concept map
○ education

Concept Map | Propositions

GoJS evaluation
(c) 1998-2016 Northwoods Software
Not for distribution or production use
nwoods.com

Copyright © PPGI - UFES 2016

Figure 8.2 CMBuilder Interface

The user actions are explained in the following sections, in sequence.

(i) **Select Representation Type:** Text can generate four different map representations, which are: Text-based Representation, Text-based Summarization, Text and Domain-based Summarization and Domain-based

Summarization. Thus, from the interface, the user can choose the representation type **(1)** suitable for his/her purpose.

(ii) **Insert Text:** From the interface, the User enters with the input text **(2)**.

(iii) **Select Domain:** The User selects the domain to represent the concept map **(3)**.

(iv) **Process Text:** From the interface, the user can activate the "Processing" option. This action must be performed after the actions Select Representation Type and Insert Text or Select Domain. For the representation of Text-based Summarization and Text and Domain-based Summarization, the CMBuilder needs the user's help to define the domain. The user chooses one of the extracted concepts to represent the domain.

After this process, the CMBuilder returns to the Main Interface a list of propositions **(5)** extracted from the text and a concept map **(4)** built automatically.

8.3 Experiments for Text Representation on Concept Map - English Language

To perform the experiments in the English language, we use as data source the Introduction Section of the article titled "*The Theory Underlying Concept Maps and How to Construct and Use Them*" (NOVAK & CAÑAS, 2008). The text is written in English and is composed of 26 sentences and 617 words.

From the article, two experiments were conducted: *(i)* Experiment for Text-based Representation, i.e., the generation of a concept map containing all identified propositions extracted from the data source and *(ii)* Experiment for Text-based Summarization, i.e., the generation of concept maps containing relevant propositions from the text.

8.3.1 Experiment for Text-based Representation

This experiment was conducted to demonstrate an overview of the features present in a concept map generated by the CMBuilder tool. The experiment identified 26 sentences, 165 propositions and 99 concepts. Figure 8.3, illustrates the output of this process without applying the Summarization step.

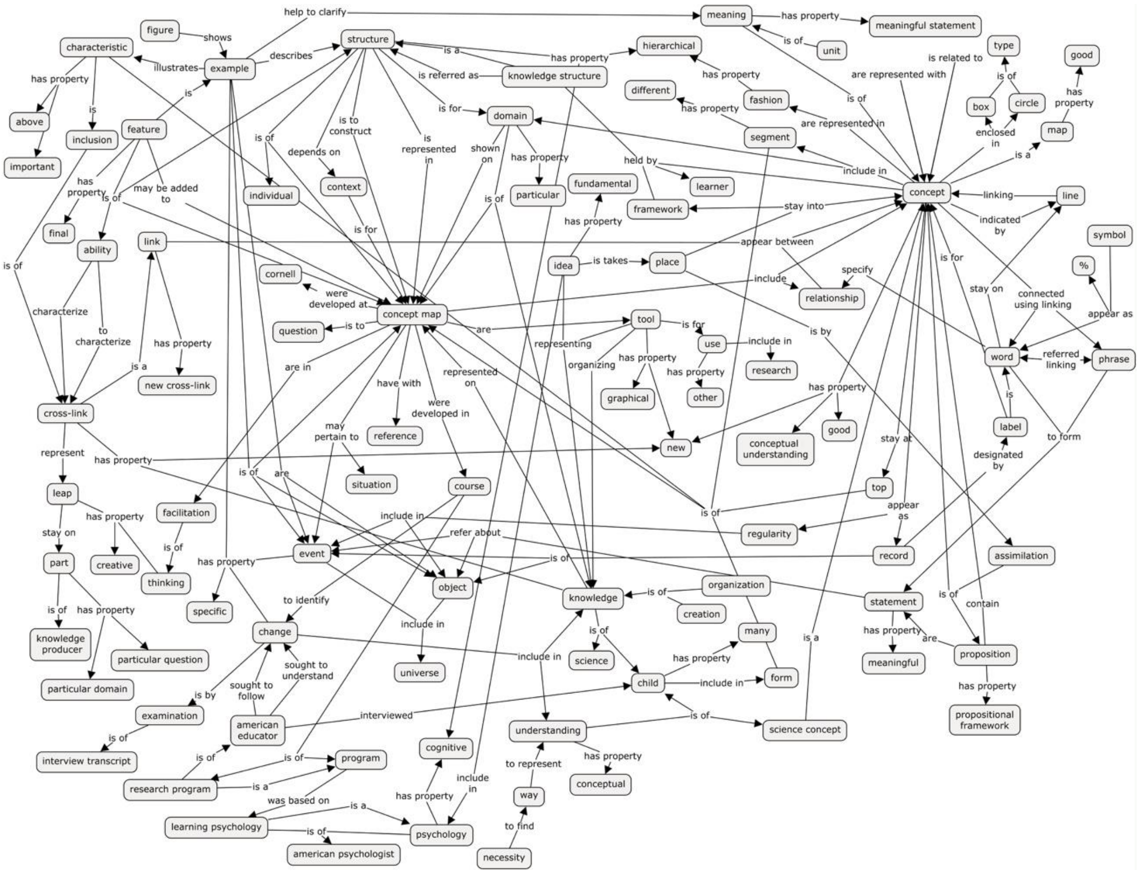


Figure 8.3 Concept map generated by the CMBuilder to English language

Here, we point out some of the features of the concept map generated by CMBuilder in this experiment:

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

(ii) **Proposition identification from Specialization relationship** - The concept "*program*" is extracted from the text "This program was based on the learning psychology...", and the concept "*research program*" is extracted from the text "...course of Novak's research program at Cornell...". The approach has created a relationship of specialization between the concept "*program*" and "*research program*", with the label "*is a*".

(iii) **Anaphora resolution** - The proposition (*concept map, include, concept*) was extracted from the text "Concept maps are graphical tools for organizing and representing

knowledge. They include concepts...". The approach associates the pronoun "*they*" to the concept "*concept map*".

(iv) **Proposition identification from distant syntactic connections** - Using the syntax tree created for the text "Figure 1 shows an example of a concept map that describes the structure of concept maps and illustrates the above characteristics.", the approach extract the distant propositions: (*figure 1, shows, example*), (*example, describes, structure*), (*example, illustrates, characteristic*), (*structure, is of, concept map*).

(v) **Similarity of concepts** - The concept "*Ausubel*", extracted from the text "The fundamental idea in Ausubel's cognitive psychology...", and the concept "*David Ausubel*", extracted from "This program was based on the learning psychology of David Ausubel...", are considered as similar concepts and are represented by the most significant label, "*David Ausubel*". The concepts "*concept*" and "*concepts*" are associated as similar concepts and represented by the label "*concept*". Our approach favors the most generic or high-level labels when there are concepts with some proximity, and more specific labels otherwise. That is, the concept "*good map*" is represented by the more general concept "*map*" and the concept "*interview transcript*" remains with its original label.

(vi) **Labeling of entities** - Concepts defined as entities of type Person are associated with their description found on DBPedia. For instance, the concept "*David Ausubel*" is associated with the URI "*American psychologist*" on DBpedia.

(vii) **Identification of multi-words concepts** - The approach adopts lexical and syntactic rules to identify more complete labels of concepts, such as "*knowledge producer*".

(viii) **Genitive interpretation** - The proposition (*research program, is of, american educator*) is extracted from "... course of Novak's research program at Cornell...". The approach identifies and transforms the genitive form into an intermediate form.

Since this experiment extracts all propositions identified from the text, no statistical analysis was performed on the result.

8.3.2 Experiment for Text-based Summarization

This experiment added the Summarization module to the process undertaken in the experiment presented in Section 8.3.1. The experiment identified 58 concepts and 121 propositions. Figure 8.4 illustrates the output of this process.

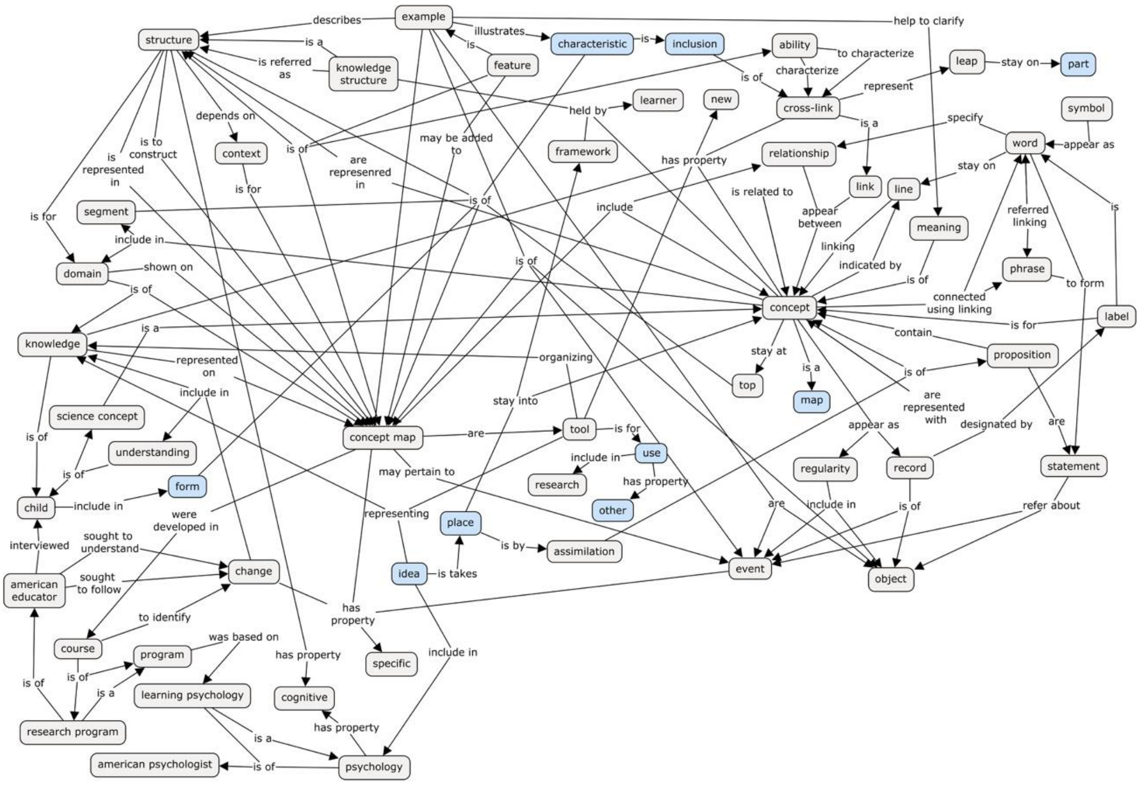


Figure 8.4 Concept map generated by CMBuilder to English language

An experimental analysis was conducted subjectively by comparing the map built by CMBuilder shown in Figure 8.6, with others from related works (see Section 4.3). Our intention is to analyze the quality of the map generated with respect to the original text.

We note some strong points associated with the map built by CMBuilder which outperformed the results reported by related works, namely: *(i)* All the concepts are connected by linking phrases without fragments. Despite the Summarization step, the resulting concept map establishes valid relationships between concepts, even for topologically distant concepts in the text; *(ii)* Concept labels are small, formed by multi-words expressions when applicable; *(iii)* Neither pronouns nor named entities make up relevant concept labels; *(iv)* Labels are directly extracted from the data source; *(v)* Relationship labels are meaningful and formed by verbs and sometimes not explicitly mentioned in the text; and (*vi)* Concepts and propositions do not exhibit any redundancy.

8.3.2.1 Analysis and Results

In order to analyze the fidelity of the generated map to the text, we compare the concept map automatically generated by CMBuilder (Figure 8.4) to concept maps manually built by ten domain experts using the same data source.

The following instructions were provided: *(i)* the experts received information about the use of concept maps in general and about the purpose of the experiment; *(ii)* they were instructed that the label of concepts and relationships should be short, meaningful and extracted from the text; *(iii)* they were informed that concepts' labels should contain nouns, and relations' labels should contain verbs; *(iv)* they were instructed that labels containing named entities or prepositions should be changed to more appropriate labels.

The following tables show the precision and recall calculated by comparing the map constructed by CMBuilder with the maps generated by the experts.

Table 8.1 shows the analysis of the identified concepts, reaching 0.75 in Precision and 0.45 in Recall. In this experiment, we disregarded the label flexion of concept maps built by experts, such as plural.

Concept Analysis

Expert	*Exp.1*	*Exp.2*	*Exp.3*	*Exp.4*	*Exp.5*	*Exp.6*	*Exp.7*	*Exp.8*	*Exp.9*	*Exp.10*	**AVG**
Precision	0.78	0.65	0.77	0.79	0.63	0.76	0.82	0.74	0.78	0.76	**0.75**
Recall	0.43	0.58	0.36	0.53	0.48	0.32	0.50	0.44	0.50	0.39	**0.45**

Table 8.1 Results for fidelity of Concepts to English language

Table 8.2 shows the analysis of the identified relationships, obtaining 0.57 in Precision and 0.23 in Recall. In this evaluation, we consider relations as similar to those generated by the experts, if they are linking the same concepts exactly and their meaning is similar.

Relationship Analysis

Expert	*Exp.1*	*Exp.2*	*Exp.3*	*Exp.4*	*Exp.5*	*Exp.6*	*Exp.7*	*Exp.8*	*Exp.9*	*Exp.10*	**AVG**
Precision	0.73	0.50	0.58	0.49	0.53	0.61	0.62	0.58	0.49	0.56	**0.57**
Recall	0.20	0.30	0.19	0.20	0.28	0.17	0.25	0.28	0.26	0.18	**0.23**

Table 8.2 Results for fidelity of Relationships to English language

The results obtained in Table 8.1 and Table 8.2 are modest mainly because of the complexity of the task, but they show promising results. Moreover, only 9 (blue color in Figure 8.4) of the 58 concepts that compose the map constructed by CMBuilder were not represented in concept maps constructed by the experts.

8.4 Experiments for Text Representation on Concept Map - Portuguese Language

To perform the experiments in the Portuguese language, we use as data source the Introduction Section of the article titled "*A teoria subjacente aos mapas conceituais e como elaborá-los e usá-los*" (NOVAK & CAÑAS, 2010). This text is similar to the text used in the Section 8.3 in English. The text is written in Portuguese and is composed of 26 sentences and 592 words.

From the article, two experiments were conducted: *(i)* Experiment for Text-based Representation, i.e., the generation of a concept map containing all identified propositions extracted from the data source and *(ii)* Experiment for Text-based Summarization, i.e., the generation of concept maps containing relevant propositions from the text.

8.4.1 Experiment for Text-based Representation

This experiment was conducted to demonstrate an overview of features present in a concept map generated by CMBuilder tool to Portuguese language. The experiment identified 26 sentences, 123 propositions and 80 concepts. Figure 8.5, illustrates the output of this process without applying the Summarization step.

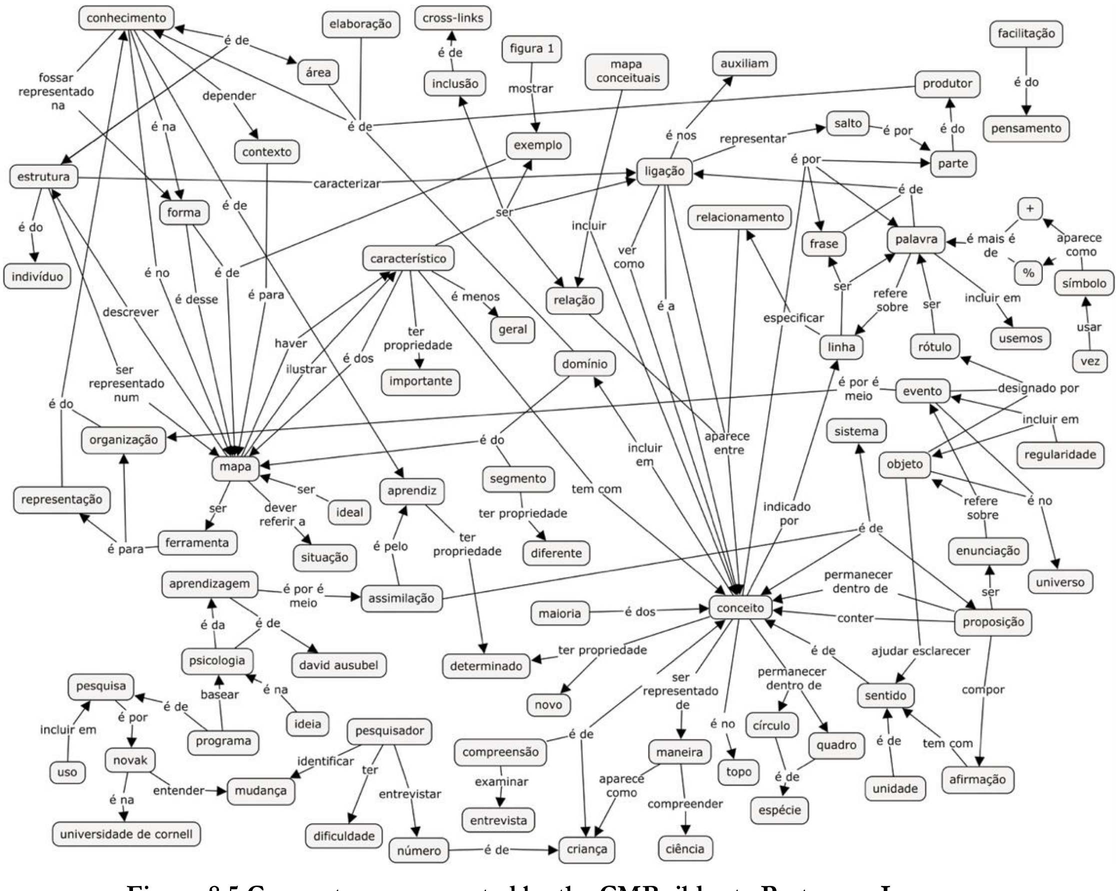


Figure 8.5 Concept map generated by the CMBuilder to Portugese Language

Here, we point out some of the features of the concept map generated by CMBuilder in this experiment:

(i) **Proposition identification from a prepositional sentence** - The proposition (*relação, aparece entre, conceito*) is extracted from the text "…que são as relações ou ligações entre conceitos nos diferentes segmentos...". The approach creates a relationship between the concepts "*relação*" and "*conceito*" with the label "*aparece*

entre". The labels are defined with the help of the prepositions mapping carried out during the Structure Adjusting step. This mapping was not very suitable to represent the relationships label in Portuguese language.

(ii) **Proposition identification from Specialization relationship** - The concepts "*aprendiz*" and "*determinado*" are extracted from the text "…conhecimento de um determinado aprendiz…". The approach has created a relationship of specialization between the concept "*aprendiz*" and "*determinado*", with the label "*ter propriedade*". Specialization relationships "*is a*" have not been extracted by the experiment.

(iii) **Anaphora resolution** - The proposition (*psicologia, é de, david ausubel*) was extracted from the text "...na psicologia cognitiva de Ausubel é que...". The approach associates the noun "*Ausubel*" to the concept "*David Ausubel*" extracted from the text "…baseava na psicologia da aprendizagem de David Ausubel…". Only anaphora containing entities were resolved by the experiment.

(iv) **Proposition identification from distant syntactic connections** - Using the syntax tree created for the text "Figure 1 shows an example of a concept map that describes the structure of concept maps and illustrates the above characteristics.", the approach extract the distant propositions: (*figura 1, mostrar, exemplo*), (*exemplo, é de, mapa*), (*mapa, descrever, estrutura*) and (*mapa, descrever, característica*).

(v) **Identification of multi-words concepts** - The approach adopts lexical and syntactic rules to identify more complete labels of concepts, such as "*mapas conceituais*".

Some of the adjustments adopted in the English language (Section 8.3) were not implemented in the Portuguese language such as Labeling of entities, Similarity of concepts and Genitive interpretation.

Many propositions were not extracted due to errors related to: *(i)* **syntactic parser**, the quality and precision of Portuguese parser is much smaller than English; *(ii)* **anaphora resolution**, only anaphora of named entities were resolved; and *(iii)* **lemmatization**, the words received incorrect lemmas, which affected the context of the proposition and damaged the identification of similar terms. Consequently, these errors caused lost of information and portions of fragmented maps.

8.4.2 Experiment for Text-based Summarization

This experiment added the Summarization module to the process undertaken in the experiment presented in Section 8.4.1. The experiment identified 53 relevant concepts and 95 related propositions. Figure 8.6 illustrates the output of this process.

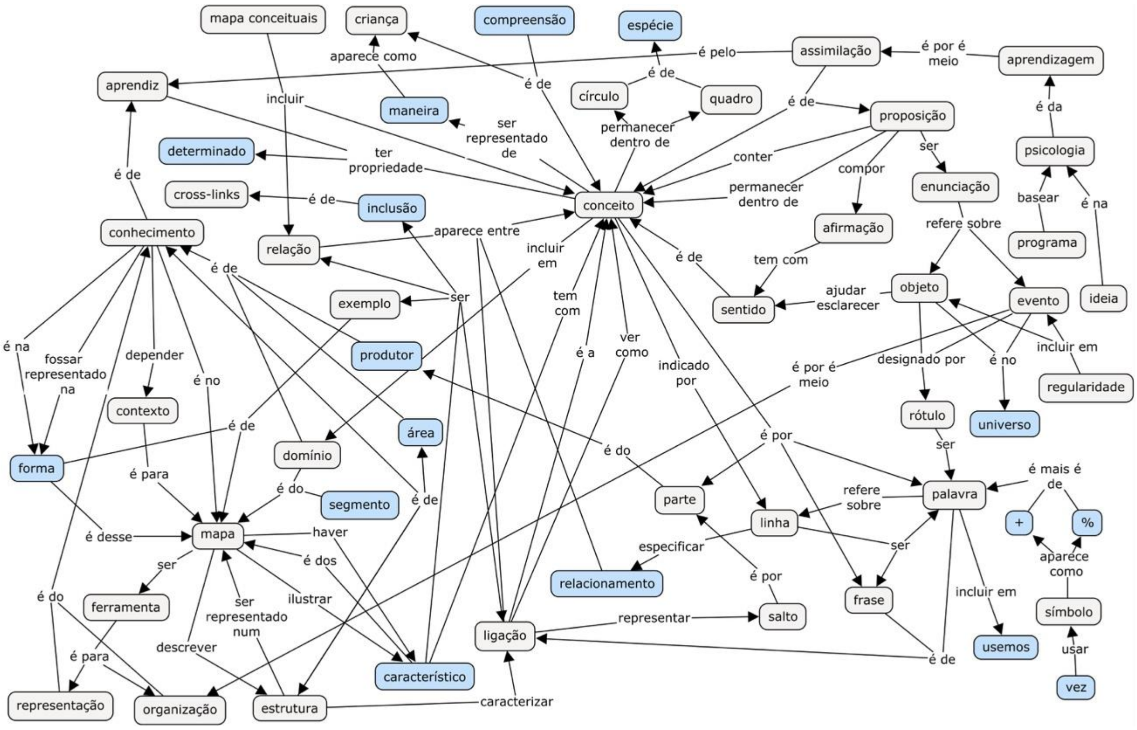


Figure 8.6 Concept map generated by CMBuilder to Portuguese language

An experimental analysis was conducted subjectively by comparing the map built by CMBuilder shown in Figure 8.6, with others from related works (see Section 4.3). The main difference of the experiment presented by CMBuilder is the use of the Portuguese language, since the related works use English, Croatian and Spanish languages (Section 4.2.3).

Besides, we note some strong points associated with the map built by CMBuilder compared with the results reported by related works, namely:

(i) All the concepts are connected by linking phrases without fragments;

(ii) Labels are directly extracted from the data source;

(iii) Concept labels are small and formed by multi-words expressions when applicable;

(iv) Relationship labels are meaningful and formed by verbs and sometimes not explicitly mentioned in the text;

(v) Concepts and propositions do not exhibit any redundancy.

And some weak points, namely:

(i) Named entities are labels to concepts;

(ii) Important information of text has been lost.

(iii) Lemmatization of concepts damaged the understanding of the propositions.

8.4.2.1 Analysis and Results

In order to analyze the fidelity of the generated map to the text, we compare the concept map automatically generated by CMBuilder (Figure 8.6) to concept maps manually built by five domain experts using the same data source.

The following instructions were provided: *(i)* the experts received information about the use of concept maps in general and about the purpose of the experiment; *(ii)* they were instructed that the label of concepts and relationships should be short, meaningful and extracted from the text; *(iii)* they were informed that concepts' labels should contain nouns, and relations' labels should contain verbs; *(iv)* they were instructed that labels containing named entities or prepositions should be changed to more appropriate labels.

The following tables show the precision and recall calculated by comparing the map constructed by CMBuilder with the maps generated by the experts. Table 8.3 shows the analysis of the identified concepts, reaching 0.68 in Precision and 0.38 in Recall. In this experiment, we disregarded the label flexion of concept maps built by experts, such as plural.

Concepts Analysis						
Expert	*Exp.1*	*Exp.2*	*Exp.3*	*Exp.4*	*Exp.5*	**AVG**
Precision	0.77	0.69	0.65	0.70	0.59	**0.68**
Recall	0.58	0.16	0.24	0.64	0.30	**0.38**

Table 8.3 Results for fidelity of Concepts to Portuguese language

Table 8.4 shows the analysis of the identified relationships, obtaining 0.41 in Precision and 0.19 in Recall. In this evaluation, we consider relations as similar to those generated by the experts, if they are linking the same concepts exactly and their meaning is similar.

Relationships Analysis						
Expert	*Exp.1*	*Exp.2*	*Exp.3*	*Exp.4*	*Exp.5*	**AVG**
Precision	0.50	0.33	0.33	0.53	0.36	**0.41**
Recall	0.29	0.05	0.08	0.41	0.11	**0.19**

Table 8.4 Results for fidelity of Relationships to Portuguese language

The low value achieved by the Recall metric can be explained by the concept maps size. Since the experts read a text in their native language and had mastery over the subject, the constructed maps were very brief and with a minimum amount of concepts.

Moreover, although the value reached by the precision and recall was not high, only 16 (blue color in Figure 8.6) of the 53 concepts that compose the map constructed by CMBuilder were not represented in concept maps constructed by the experts.

8.5 Research on the Manual Construction of Concept Maps from Texts

In order to understand the process and difficulties of the manual construction of concept maps from texts, we conducted a survey with 10 experts in the domain of concept maps whose native language is not English.

Data collection was conducted through a questionnaire containing 12 closed and open questions. Appendix I shows the questionnaire that aims to identify the difficulties in the process of manual construction of concept maps from texts.

The survey was conducted as follows: *(i)* The experts received information about the use of concept maps in general and about the purpose of the research; *(ii)* They received a text in English containing 630 words, which is the same that was applied in the experiment performed in Section 8.3; *(iii)* They were instructed to construct a concept map of essentially scientific nature from that text, i.e., concepts' labels should contain nouns, and relations' labels should contain verbs; *(iv)* After the manual construction of the concept map, the experts were instructed to answer a questionnaire.

From the experiment and questionnaire answered by the experts, we can collect and highlight some information. Figure 8.7 shows the time taken by experts to build the concept map from the text. As shown by the graph, the average time to construct manually the concept map is 1 hour and 47 minutes.

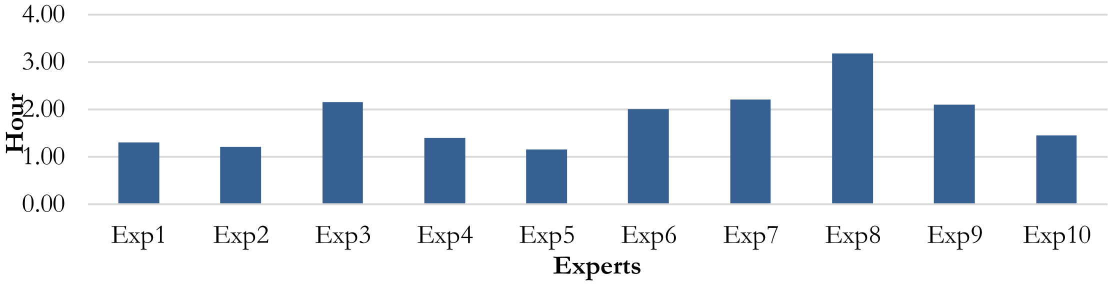


Figure 8.7 Time taken by expert

Figure 8.8 shows the level of ease identified by the specialist to construct the map. As shown by the graph, the task of building concept map from text was considered at the average level of ease 4.5.

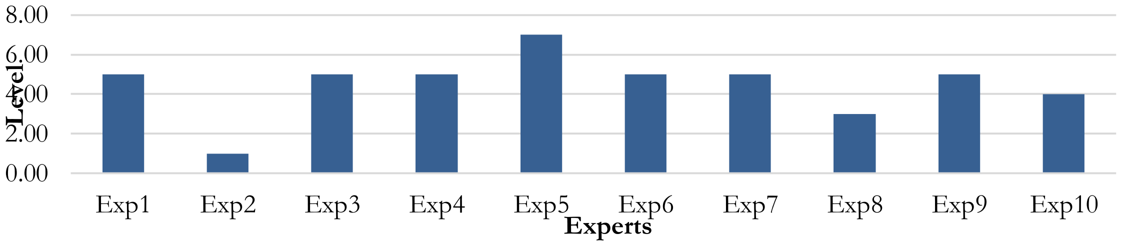


Figure 8.8 Level of ease to construct concept map from text

8.5.1 Analysis and Results

In analyzing the data, we see some important information that guide the development of this research, as follows:

(i) The average time for manual construction of a concept map from a text containing about 630 words is greater than 1hr.

(ii) The greatest difficulty during the construction of the map is related to the identification of the relations (100%) and, in some cases, the author cannot represent all the concepts (62.5%) and relationships (75%) that he/she considers relevant.

(iii) The authors (100%) consider that this task requires high cognitive effort and an intermediate skill level with English. Most of them (60%) consider that this task has a degree of ease 5, in a scale from 1 to 10.

(iv) The authors classified this activity as tiring (75%), motivating (75%), stimulating (37.5%) and stressful (25%).

(v) The task helps text understanding (75%), since the author must *(i)* read and reread the text to extract concepts, *(ii)* to deepen the understanding of the text to extract relationships, and *(iii)* to find a proper way to represent and connect the concepts in the map.

(vi) Although some authors (25%) consider that the use of a technological tool for this purpose could lead to losses in the understanding process, all authors (100%) agree that the tool would bring great benefits to the text understanding.

We conclude that CMBuilder would be very useful because it would reduce time taken for the construction of a concept map. Besides it would enable the construction of knowledge structures from complex and unknown texts. Therefore, the tool is not directed to the construction of a concept map but to allow a reflection, an analysis and a review of the map, and an observation of the concepts and their interconnections. In other words, the tool is being used mainly as a support for understanding text and knowledge construction as well.

8.6 Research on the Influence that Summarized Concept Map has for the Understanding of the Subject

In order to understand the influence that the concept map automatically summarized from a text has for the understanding of the subject addressed in that text, we conducted an experiment with 12 master's degree students in Computer Science of Federal University of Espírito Santo. Since the experiment dealt with known subjects, students were instructed not to take advantage of their prior knowledge but to stick only to the information presented.

The data collection was conducted by means of a questionnaire prepared with the following resources: *Text I* containing 123 words and *Text II* containing 302 words. These texts were submitted to CMBuilder that summarized them in *Map I* and *Map II*.

Two questionnaires, *A* and *B*, were prepared from those resources, each one composed of two steps: *Step 1*, analysis from multiple-choice questions, and *Step 2*, analysis from discursive question.

(i) **Quiz A** (Appendix B):

- Step 1: *Text I* with 5 multiple choice questions for text comprehension;
- Step 2: *Map II* with a discursive question for map comprehension.

(ii) **Quiz B** (Appendix C)

- Step 1: *Map I* with 5 multiple choice questions for map comprehension;
- Step 2: *Text II* with a discursive question for text comprehension.

The students were organized into two groups, A and B, respectively receiving questionnaires A and B. The experiment aimed to compare the answers of the groups to analyze the information extracted and assimilated from the text and the map.

8.6.1 Analysis and Results

From the questionnaire answered by the students, we can collect and highlight some information. Figure 8.9 shows the score achieved by *Groups A* and *B* during *Step 1*. As shown by the graph, the *Group A*, using text, had higher score (97%) than *Group B*, using map (60%).

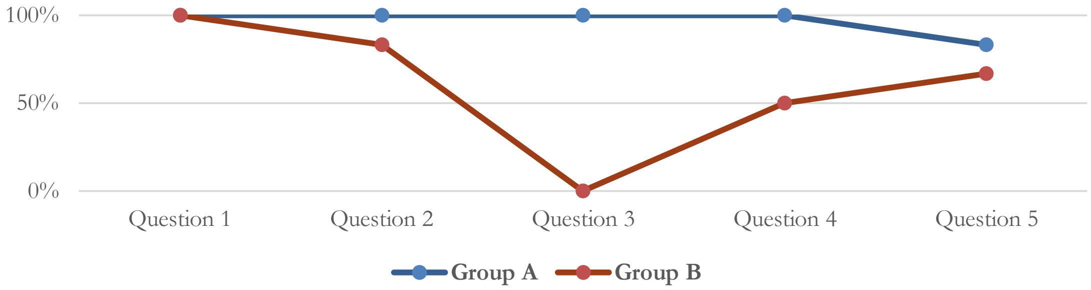


Figure 8.9 Score achieved by groups *A* and *B* during *Step 1*

This score shows the summarized map was not able to cover all the information requested in the questionnaire. Only the information presented in questions 1 and 2 were clear on the map.

Figure 8.10 shows the scores achieved by students in Groups A and B during Stage 2. Since Step 2 is composed of one discursive question, we use the following contents distribution to analyze the answers: What concept map is? *(20%)*; What is it composed of? *(20%);* Where did the main idea come from? *(20%);* Who created it? *(20%);* besides of the presence of following relevant concepts: concept map, tool, representation, concept, relation, Novak, Ausubel, knowledge, psychology, research and child *(20%).*

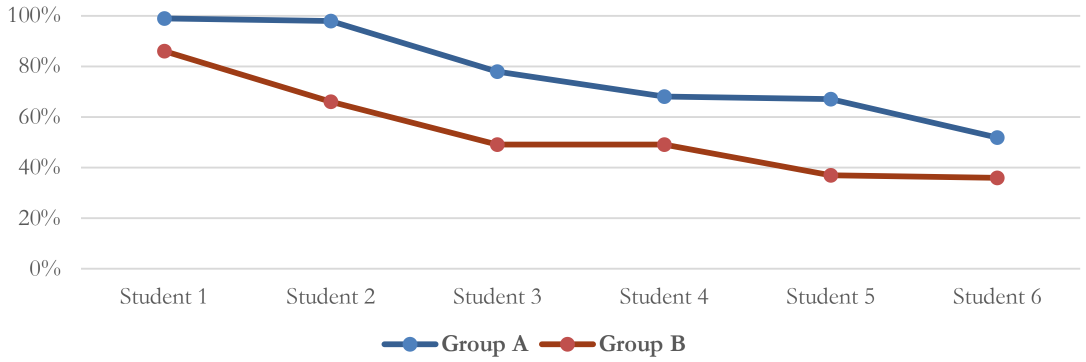


Figure 8.10 Score achieved by groups *A* and *B* during *Step 2*

As shown by the graph, the *Group A*, using map, had higher results (77%) than *Group B*, using text (54%). This result is justified by the summarized concept map presenting sufficient and objective information for the interpretation of the subject addressed. The lower score of *Group B* comes from the fact that students were unable to extract all relevant information from the text, since the text is grammatically complex and extensive.

We verified a great difference in the answers elaborated by the students of *Group A* and *Group B*. *Group A*, using map, elaborated more objective and complete answers while *Group B*, using text, elaborated more dispersed and non-objective answers, as exemplified in Table 8.5. The table presents the highest and lowest scores achieved by each group, whose relevant concepts are highlighted in *italic*.

Answers Analysis

Group	Answer	Score
A	Um *mapa conceitual* é uma *ferramenta* utilizada na *psicologia* para o *aprendizado*. Ajuda no aprendizado de *crianças*. Foi desenvolvido por *Novak* baseado nas teorias cognitivas de *Ausubel*. Inclui *relações* e *conceitos* que podem ser palavras, frases ou objetos diversos para *representar* um *conhecimento* específico.	99%
A	*Mapas conceituais* são *ferramentas* de natureza *cognitiva*, desenvolvidas ao longo de um programa de *pesquisa*, utilizadas para *representar conhecimento*. Se baseia na *psicologia*, proposta por uma pessoa chamada *Ausubel* e gira em volta do termo "*conceito*", indicando que a estrutura dos mapas conceituais existem para representar esses conceitos e torná-los legíveis para quem se propõe a estudá-los.	52%
B	*Mapas conceituais* são representações *gráficas* de *conceitos* e suas *relações*. Nos mapas conceituais, os conceitos são representados por caixas ou círculos contendo uma ou mais palavras, e suas relações são dadas por linhas ligando duas caixas. Foram propostos por *Novak*, baseado nas	86%

	ideias de *Ausubel*, com o objetivo inicial de *estudar* as mudanças no mapa de *conhecimento* de *crianças*.	
B	O texto fala sobre os *mapas conceituais*, como são suas representações e como os mesmos são apresentados com suas *ligações* e o que significam. O texto também faz menção a um pouco da história dos mapas conceituais, sua criação e como os mesmos podem gerar *conhecimento* específico sobre variados assuntos.	**36%**

Table 8.5 Sample of the highest and lowest scores achieved by each group

Although the research is limited, it gives evidence that the use of a concept map summarized from text can contribute to the construction of knowledge. We emphasize that its success depends on the quality of the concept map used. Moreover, we can consider that a text, especially extensive, presents great difficulties to the reader to identify relevant information and consequently assimilates them.

8.7 Some Considerations on the Chapter

This chapter presented the CMBuilder and some aspects of its development, operation and execution. In addition, we apply experiments to validate the developed tool and the use of a concept map for text summarization.

Through the experiments we can highlight some important points, as follows: *(i)* The CMBuilder was able to construct an concept map of scientific style from the text, reaching 0.75 precision and 0.45 recall for concepts and 0.57 precision and 0.23 recall for relationships in English language, and reaching 0.68 precision and 0.38 recall for concepts and 0.41 precision and 0.19 recall for relationships in Portuguese language; *(ii)* The manual construction of concept maps requires great effort and time, since the same text was constructed in less than 1 minute by the tool and more than 1 hour by the expert.

From the experiment, we note that building a map from a text is a difficult task, even for domain experts. In fact, the experts were told not to represent the concepts in their cognitive structure. Instead, they were instructed to use only the concepts expressed in the text, which they found difficult. Moreover, experts did not construct their best concept maps, because the text used in the experiment was considered extensive. This fact caused demotivation, increased cognitive effort and time spent. After 30 min of experiment, the experts drastically reduced the quality of the map being built.

By means of observations we noticed some advantages in using the proposed tool, especially compared to related work. Nevertheless, we still have other challenges that can be summarized as follows:

(i) The anaphora resolution is still far from satisfactory, especially with respect to demonstrative and possessive pronouns.

(ii) Some assigned labels do not correspond to the labels assigned by the experts. The CMBuilder, sometimes, did not make use of some adjectives and adverbs relatively important for characterizing the labels.

(iii) Some relationships assigned by the experts were not explicitly extracted from the text because the pre-existing information in their cognitive structure interfered in their representation of the map. Thus, it was not possible to directly compare them to our extracted relations.

(iv) Some relevant domain concepts were lost during the Summarization module.

(v) The text used to extract propositions and construct the map must be scientific-style and contain concise information, i.e., it does not process any text type.

Finally, we conducted an experiment to verify if the concept map summarized by CMBuilder has influence for the understanding of the subject addressed in a text. The experiment has shown that the use of the maps is satisfactory since it reached 60% of hits for maps extracted from small texts with multi-choice questions and 77% of hits for maps extracted from extensive texts with discursive questions. Although the experiment shows evidence of the validity of the summarized map for the construction of knowledge, it has little value due the number of students involved.

The next chapter presents and discusses the research conclusions, as well as the future works.

Chapter 9
Final Considerations and Future Work

This chapter presents some considerations of all the work developed by this research, as well as of the selection of future works essential for the continuity and improvement of this research.

This chapter is organized as follows: Section 9.1 presents the final considerations; and Section 9.2 discuss future works.

9.1 Final Considerations

In order to answer or validate the hypotheses presented in Chapter 1, this research started with a literature review on the technological approaches directed to the construction of concept maps. The literature review resulted in a categorization to better identify and analyze the functionalities and characteristics of the technological approaches in this context.

The categorization was used to visualize and analyze comprehensively and accurately the main features adopted in each approach and served as the basis for the definition of our conceptual model. From the conceptual model, we noticed that none of the related works included the characteristics adopted by CMBuilder following the categorization, since it combines: *(i)* domain identified; *(ii)* linguistic manipulation method; *(iii)* its own interface; and *(iv)* an automatic process.

From the conceptual model, we defined a technological architecture to satisfy the objective of this research. This computational architecture was applied to the development of CMBuilder tool, which is publicly available.

From the experiments performed with the CMBuilder tool, we can observe the quality of the concept map built. Compared with maps constructed by other approaches, we conclude that the map constructed by CMBuilder includes important characteristics that make it an acceptable representation for a text and superior to other related approaches. On the maps constructed by experts, we conclude that they maintain the author's individual view despite being a representation of the text. This prevents an objective and analytical analysis, interfering with the results of the experiment.

Besides the points presented, we emphasize that no related work is publicly available for use, download, extension or service. In this context, CMBuilder tool brings a great contribution to Education and Research. Although the main objective of the research is the CMBuilder tool, this research also enabled the development of ExtroutNLP API, an API composed of several information extraction libraries available for use and extension by other projects.

From the experiments performed with the ExtroutNLP API we can observe some characteristics of the models and libraries developed. Compared with other libraries, we conclude that ExtroutNLP:

(i) Extracts different patterns of triples resembling to propositions definition;

(ii) Maintains order of the ranking similar to other approaches; and

(iii) Adopts a specialized method for summarization of propositions;

(iv) Works with English and Portuguese language.

Therefore, considering the research developed in the course of this dissertation, we can conclude that the following hypotheses were validated:

(i) Based on Chapter 8, we conclude that it is possible to create a public tool to automatically construct concept maps from the texts. Looking at the experiments presented in Sections 8.3.1 and 8.3.2, we conclude that the concepts maps are of scientific style and it can represent a summarization of a text.

(ii) Based on the technological architecture presented in Chapter 6, we can conclude that the variation of linguistic components (tokenizer and parser) can provide the research expansion for multilingual application in Portuguese and English languages. Based on the experiments presented in Section 7.3.4, we can conclude that techniques based on linguistic structure are relatively competent to extract propositions from texts.

(iii) Based on the experiments presented in Section 8.6, we can conclude that the use of a concept map summarized from the text influences the process of understanding the text self.

However, the following hypotheses have not been validated to date:

(i) By using the domain identifier module, we cannot conclude that the use of a domain knowledge base impact the quality of the concept map.

Finally, we can consider that works dedicated to the automatic construction of concept maps are relatively new and still evolving. The CMBuilder has shown promising results, although some challenges are not yet satisfactorily resolved.

9.2 Future Works

Future works will be focused on the quality of CMBuilder tool, development of libraries for ExtroutNLP API and the studies on the summarization of concept maps.

Since CMBuilder tool is available publicly, we need to devote our efforts to ensure the quality of the developed features. Therefore, our future works will be related to testing and improving of the functionality, usability and efficiency of the CMBuilder tool.

Future works on ExtroutNLP can not be readily estimated, since API intends to gather a set of solutions for information extraction from texts in Portuguese and English. The following outlines some of the future projects in this regard.

For the OpenIE library we emphasize the: *(i)* Improvement of the anaphora resolution process; *(ii)* The adoption of a disambiguation layer between the library and DBPedia ensuring a more appropriate concepts labeling; *(iii)* The identification of apposed and hyponyms; *(iv)* The study of different approaches for a better propositions extraction; *(v)* The consideration of multiword expression in defining concepts and relationships; *(vi)* Of the use of a semantic network for the identification of relations; *(vii)* The improvement the accuracy of the parser for the Portuguese language.

For the Summarization library we stress the importance of studies related to the use of thesaurus to the ranking, as well as to identify the best parameters. Besides we stress the importance of the works directed to tests and validation of the proposed method in different corpora.

In addition, we stress the importance of the work on the summarization of concept maps. However, apart from this dissertation, more accurate and in-depth studies on the impact of concept maps summarized from texts for text understanding need to be conducted. This could be a great contribution to education in general.

Finally, for broader and more accurate results, we will soon make available the tool for teachers and students of the state's public network.

References

AFONSO, S., BICK, E., HABER, R. & SANTOS, D., 2002. Floresta Sintá (c) tica: A treebank for Portuguese.

AGUIAR, C. Z. & CURY, D., 2016. *A categorization of technological approaches to concept maps construction.* Costa Rica, Learning Objects and Technology (LACLO), pp. 1-9.

AGUIAR, C. Z., CURY, D. & ZOUAQ, A., 2016. *Automatic Construction of Concept Maps from Texts.* s.l., s.n.

AITCHISON, J., GILCHRIST, A. & BAWDEN, D., 2000. *Thesaurus construction and use: a practical manual.* s.l.:Psychology Press.

AJLI, A. & AFDEL, K., 2014. *A new hybrid approach for constructing the concept map based on fuzzy prerequisite relationships.* s.l., s.n., pp. 115-121.

AL-SAREM, M., BELLAFKIH, M. & RAMDENI, M., 2011. An approach for mining concepts' relationships based on historical assessment records.. *Procedia Engineering,* pp. 3245-3249.

ALUISIO, S. et al., 2003. *An account of the challenge of tagging a reference corpus for brazilian portuguese.* s.l., s.n.

ANDER-EGG, E., 1978. Introducción a las técnicas de investigación social para trabajadores sociales.

ARAMPATZIS, A. T., VAN DER WEIDE, T. P., VAN BOMMEL, P. & KOSTER, C. H., 1999. Linguistically-motivated information retrieval.

AUER, S. et al., 2007. *Dbpedia: A nucleus for a web of open data.* s.l., Springer Berlin Heidelberg.

AUSUBEL, D. P., NOVAK, J. D. & HANESIAN, H., 1968. Educational psychology: A cognitive view. pp. 15-31.

BAEZA-YATES, R. & RIBEIRO-NETO, B., 2013. *Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca.* 2 ed. s.l.:Bookman.

BAI, S.-M. & CHEN, S.-M., 2008. *A new method for automatically constructing concept maps based on data mining techniques.* s.l., IEEE, pp. 3078-3083.

BAI, S.-M. & CHEN, S.-M., 2008. Automatically constructing concept maps based on fuzzy rules for adapting learning systems. *Expert systems with Applications,* Volume 35, pp. 41-49.

BANKO, M. et al., 2007. *Open Information Extraction from the Web.* s.l., s.n., pp. 2670-2676.

BÄR, D., ZESCH, T. & GUREVYCH, I., 2013. *DKPro Similarity: An Open Source Framework for Text Similarity.* s.l., s.n., pp. 121-126.

BIBER, D., CONRAD, S. & REPPEN, R., 1998. *Corpus linguistics: Investigating language structure and use.* s.l.:Cambridge University Press.

BICHINDARITZ, I. & AKKINENI, S., 2006. Concept mining for indexing medical literature.. *Engineering Applications of Artificial Intelligence,* Volume 19, pp. 411-417.

BIRD, S., KLEIN, E. & LOPER, E., 2009. *Natural language processing with Python.* s.l.:O'Reilly Media, Inc..

BRANCO, A. et al., 2010. *Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: the CINTIL DeepGramBank.* s.l., s.n.

CAÑAS, A. J. et al., 2003. A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support.. *Pensacola.*

CHARNIAK, E. & MCDERMOTT, D., 1998. *Introduction to Artificial Intelligence.* s.l.:Pearson.

CHEN, D. & MANNING, C. D., 2014. *A Fast and Accurate Dependency Parser using Neural Networks.* s.l., s.n., pp. 740-750.

CHEN, N.-S., KINSHUK, P., WEI, C. W. & CHEN, H. J., 2006. *Mining e-learning domain concept map from academic articles..* s.l., IEEE, pp. 694-698.

CHEN, N.-S., WEI, C.-W. & CHEN, H.-J., 2008. Mining e-Learning domain concept map from academic articles. *Computers & Education ,* Volume 50, pp. 1009-1021.

CHEN, S.-M. & SUE, P.-J., 2013. Constructing concept maps for adaptive learning systems based on data mining techniques. *Expert Systems with Applications,* p. Expert Systems with Applications.

CLARIANA, R. B. & KOUL, R., 2004. *A computer- based approach for translating text into concept map-like representations.* s.l., s.n., pp. 14-17.

CORRÊA, A. C. G., 2003. *Recuperação de Documentos baseada em Informação Semântica no ambiente AMMO.* s.l.:Dissertação de Mestrado em Ciência da Computação UFSCAR.

COVER, T. & HART, P., 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory,* Volume 4, pp. 21-27.

COWIE, J. & LEHNERT, W., 1996. Information extraction. *Communications of the ACM,* Volume 39, pp. 80-91.

CURY, D., PERIN, W. & SANTOS JR, P. S., 2014. *CMPaaS–A platform of services for construction and handling of concept maps.* s.l., s.n.

DE LA VILLA, M., APARICIO, F., MAÑA, M. J. & DE BUENAGA, M., 2012. *A learning support tool with clinical cases based on concept maps and medical entity recognition.*. s.l., ACM, pp. 61-70.

DE OLIVEIRA, E. R. H. N. M. A. B. H. G. &. C. P. M., 2015. *Using the cluster-based tree structure of k-nearest neighbor to reduce the effort required to classify unlabeled large datasets.*. s.l., IEEE.

DEL CORRO, L. & GEMULLA, R., 2013. *Clausie: clause-based open information extraction.* s.l., ACM, pp. 355-366.

DHURIA, S., 2015. Natural Language Processing: An approach to Parsing and Semantic Analysis. *International Journal of New Innovations in Engineering and Technology.*

DODDINGTON, G. R. et al., 2004. *The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation.* s.l., s.n.

EBECKEN, N. F., LOPES, M. C. S. & COSTA, M. C., 2003. *Mineração de textos. Sistemas inteligentes: fundamentos e aplicações.* s.l.:s.n.

EICHLER, K., HEMSEN, H. & NEUMANN, G., 2008. *Unsupervised Relation Extraction From Web Documents.* s.l., s.n.

ELHOSEINY, M. & ELGAMMAL, A., 2012. *English2mindmap: An automated system for mindmap generation from english text.* s.l., IEEE, pp. 326-331.

ETZIONI, O. et al., 2011. *Open Information Extraction: The Second Generation.* s.l., s.n., pp. 3-10.

ŽUBRINIĆ, K., OBRADOVIĆ, I. & SJEKAVICA, T., 2015. *Implementation of method for generating concept map from unstructured text in the Croatian language.* s.l., IEEE, pp. 220-223.

FADER, A., SODERLAND, S. & ETZIONI, O., 2011. *Identifying relations for open information extraction.* s.l., Association for Computational Linguistics, pp. 1535-1545.

FARUQUI, M. & KUMAR, S., 2015. Multilingual open relation extraction using cross-lingual projection.

FELDMAN, R. & SANGER, J., 2007. *The text mining handbook: advanced approaches in analyzing unstructured data.* s.l.:Cambridge University Press.

FELLBAUM, C., 1998. *WordNet.* s.l.:Blackwell Publishing Ltd.

FINKEL, J. R., GRENAGER, T. & MANNING, C., 2005. *Incorporating non-local information into information extraction systems by gibbs sampling.* s.l., Association for Computational Linguistics, pp. 363-370.

FONSECA, E. R. & ROSA, J. L. G., 2013. *Mac-Morpho revisited: Towards robust part-of-speech tagging.* s.l., s.n., pp. 98-107.

FOSKETT, D. J., 1997. *Thesaurus.* s.l., Morgan Kaufmann Publishers Inc, pp. 111-134.

FRAKES, W. B. & BAEZA-YATES, R., 1992. Information retrieval: data structures and algorithms.

GAINES, B. R. & SHAW, M. L., 1994. Using knowledge acquisition and representation tools to support scientific communities. *AAAI,* pp. 707-714.

GAMALLO, P., GARCIA, M. & FERNÁNDEZ-LANZA, S., 2012. *Dependency-based open information extraction.* s.l., Association for Computational Linguistics, pp. 10-18.

GASPERIN, C. V. & LIMA, V. L. S., 2000. *Fundamentos do processamento estatístico da linguagem natural.* s.l.:PUC-RS.

GAVA, T. B. S., MENEZES, C. d. & CURY, D., 2003. *Aplicações de mapas conceituais na educação como ferramenta metacognitiva.* s.l., s.n.

GIL, A. C., 2008. *Métodos e técnicas de pesquisa social.* Sao Paulo: Atlas.

GRAUDINA, V. & GRUNDSPENKIS, J., 2008. *Concept map generation from OWL ontologies.* Finland, s.n.

HAHN, U. & MANI, I., 2000. The challenges of automatic summarization. *Computer,* Volume 33, pp. 29-36.

HASAN, R. & HALLIDAY, M. A., 1976. *Cohesion in English.* s.l.:London: Longman.

HOBBS, J. R. et al., 1997. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. *Finite-State Language Processing.*

HUDDLESTON, R., 1984. *Introduction to the Grammar of English.* s.l.:Cambridge University Press.

HUTCHINS, J., 1987. Summarization: Some problems and methods. *Meaning: The frontier of informatics,* Volume 9, pp. 151-173.

JACOB, E. K., 2004. Classification and categorization: a difference that makes a difference. *Library trends,* Volume 52, p. 515.

KANTARDZIC, M., 2011. *Data mining: concepts, models, methods, and algorithms.* s.l.:John Wiley & Sons.

KARANNAGODA, E. L. et al., 2013. *Document analysis based automatic concept map generation for enterprises..* s.l., IEE, pp. 154-159.

KHOO, C. S. & NA, J.-C., 2006. Semantic relations in information science. *Annual review of information science and technology,* Volume 40.

KLEINBERG, J. M., 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM),* Volume 46, pp. 604-632.

KLEINBERG, J. M., 1999. Authoritative sources in a hyperlinked environment.. *Journal of the ACM (JACM),* Volume 46, pp. 604-632.

KLEIN, D. & MANNING, C. D., 2003. *Accurate unlexicalized parsing.* s.l., Association for Computational Linguistics, pp. 423-430.

KODRATOFF, Y., 1999. Knowledge discovery in texts: A definition and applica- tions. Lecture. *Lecture Notes in Computer Science,* p. 16–29.

KUMAZAWA, T. et al., 2009. Toward knowledge structuring of sustainability science based on ontology engineering. *Sustainability Science,* pp. 99-116.

LAU, R. et al., 2009. Toward a fuzzy domain ontology extraction method for adaptive e-learning. *IEEE transactions on knowledge and data engineering,* pp. 800-813.

LAU, R. Y., CHUNG, A. Y., SONG, D. & HUANG, Q., 2007. *Towards fuzzy domain ontology based concept map generation for e-learning.* s.l., Springer Berlin Heidelberg, pp. 90-101.

LE COADIC, Y.-F., 1996. *A ciência da informação.* s.l.:Briquet de lemos Livros.

LEAKE, D., MAGUITMAN, A. & REICHHERZER, T., 2004. *Understanding knowledge models: Modeling assessment of concept importance in concept maps.* s.l., s.n.

LEE, C.-H., LEE, G.-G. & LEU, Y., 2009. Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning. *Expert Systems with Applications,* pp. 1675-1684.

LEE, H. et al., 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics,* pp. 885-916.

LEE, J. H. & SEGEV, A., 2012. Knowledge maps for e-learning. *Computers & Education,* pp. 353-364.

LEE, L.-Y., LIN, Y.-S. & CHU, C.-P., 2012. *Enhancement of personal concept map constructing for effective assessment.* s.l., s.n., pp. W1A-1-W1A-7.

LEE, S., PARK, Y. & YOON, W. C., 2015. Burst analysis for automatic concept map creation with a single document. *Expert Systems With Applications,* Volume 42, pp. 8817-8829.

LI, G. et al., 2008. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. *ACM SIGMOD international conference on Management of data,* pp. 903-914.

LI, H., BOLLEGALA, D., MATSUO, Y. & ISHIZUKA, M., 2011. *Using graph based method to improve bootstrapping relation extraction.* s.l., Springer Berlin Heidelberg.

LIN, D., 1998. *Extracting collocations from text corpora.* s.l., s.n., pp. 57-63.

LIPIZZI, C., DESSAVRE, D. G., LANDOLI, L. & MARQUES, J. E. R., 2016. Towards computational discourse analysis: A methodology for mining Twitter backchanneling conversations. *Computers in Human Behavior,* pp. 782-792.

LUHN, H. P., 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development,* Volume 1, pp. 309-317.

LYONS, C., 1986. The syntax of English genitive constructions. *Journal of Linguistics,* Volume 22, pp. 123-143.

MACQUEEN, J., 1967. *Some methods for classification and analysis of multivariate observations..* s.l., s.n., p. 281–297.

MAGGIORE, F. & ANZALDI, C., 1998. Interactive thesaurus construction methods in the ecological domain. *Coenoses,* Volume 13, pp. 89-98.

MANNING, C. D. et al., 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *ACL (System Demonstrations) ,* pp. 55-60.

MARCONI, M. A. & LAKATOS, E. M., 2004. *Metodologia Científica.* 4 ed. Sao Paulo: Atlas.

MATTMANN, C. & ZITTING, J., 2011. *Tika in action.* s.l.:Manning Publications Co..

MCBRIDE, B., 2001. *Jena: Implementing the rdf model and syntax specification.* s.l., CEUR-WS. org, pp. 23-28.

MCGARRY, K. & DE LEMOS, H. V., 1999. *O contexto dinânico da informação: uma análise introdutória.* s.l.:Briquet de Lemos.

MIHALCEA, R., CORLEY, C. & STRAPPARAVA, C., 2006. *Corpus-based and knowledge-based measures of text semantic similarity.* s.l., s.n., pp. 775-780.

MILLER, G. A., 2005. *WordNet: a lexical database for English.* s.l., s.n., pp. 39-41.

MITKOV, R., 2014. *Anaphora resolution.* s.l.:Routledge.

MOONEY, R. J. & BUNESCU, R., 2005. Mining knowledge from text using information extraction. *ACM SIGKDD explorations newsletter,* Volume 7, pp. 3-10.

NADEAU, D. & SEKINE, S., 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes,* Volume 30, pp. 3-26.

NGUYEN, T. H., PLANK, B. & GRISHMAN, R., 2015. Semantic Representations for Domain Adaptation: A Case Study on the Tree Kernel-based Method for Relation Extraction. *ACL,* pp. 635-644.

NONAKA, I. & TAKEUCHI, H., 1997. *Criação de conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação.* Rio de Janeiro: Campus.

NOVAK, J. D. & CAÑAS, A. J., 2008. *The theory underlying concept maps and how to construct and use them,* s.l.: s.n.

NOVAK, J. D. & CAÑAS, A. J., 2010. A teoria subjacente aos mapas conceituais e como elaborá-los e usá-los. *Práxis Educativa,* Volume 5, pp. 9-29.

OLNEY, A., CADE, W. & WILLIAMS, C., 2011. *Generating concept map exercises from textbooks.* s.l., Association for Computational Linguistics, pp. 111-119.

PÉREZ, C. C. C. & VIEIRA, R., 2005. *Mapas Conceituais: geração e avaliação.* s.l., s.n., pp. 2158-2167.

PAGE, L., BRIN, S., MOTWANI, R. & WINOGRAD, T., 1999. The PageRank citation ranking: bringing order to the web.

PEDERSEN, T., PATWARDHAN, S. & MICHELIZZI, J., 2004. *WordNet:: Similarity: measuring the relatedness of concepts.* s.l., Association for Computational Linguistics, pp. 38-41.

PERIN, W. A., CURY, D. & MENEZES, C. S., 2014. *NLP-Imap: Integrated solution based on question-answer model in natural language for an inference mechanism in concepts maps.* s.l., s.n.

PETERSEN, K., FELDT, R., MUJTABA, S. & MATTSSON, M., 2008. *Systematic Mapping Studies in Software Engineering.* s.l., s.n., pp. 68-77.

PINOTTE, G. N., CURY, D. & ZOUAQ, A., 2015. *ONTOMAP: From Concept Maps to Shallow OWL Ontologies.* s.l., s.n.

PIPITONE, A., CANNELLA, V. & PIRRONE, R., 2014. Automatic concept maps generation in support of educational processes.. *Journal of e-Learning and Knowledge Society,* Volume 10.

PIRNAY-DUMMER, P. & IFENTHALER, D., 2011. Reading guided by automated graphical representations: How model-based text visualizations facilitate learning in reading comprehension tasks. *Instructional Science,* Volume 39, pp. 901-919.

POLETTINI, N., 2004. The vector space model in information retrieval-term weighting problem. *Entropy,* pp. 1-9.

POROSHIN, V. A., 2014. *Semantic analysis of Natural Language.* s.l., s.n., pp. 16-23.

PRETI, D., 2006. *Fala e escrita em questão.* s.l.:Editora Humanitas.

QASIM, I., JEONG, J. W., HEU, J. U. & LEE, D. H., 2013. Concept map construction from text documents using affinity propagation. *Journal of Information Science,* pp. 719-736.

QUIVY, R. & CAMPENHOUDT, L. V., 2005. *Manual de Investigação em Ciencias Sociais.* 4 ed. Lisboa: Gravida.

REICHHERZER, T. & LEAKE, D., 2006. *Understanding the role of structure in concept maps.* s.l., s.n., pp. 2004-2009.

REZENDE, S. O., 2003. *Sistemas inteligentes: fundamentos e aplicações.* s.l.:Editora Manole Ltda.

RICHARDSON, R. & FOX, E., 2005. *Using concept maps in digital libraries as a cross-language resource discovery tool.* s.l., ACM, pp. 256-257.

SALTON, G. & BUCKLEY, C., 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management,* Volume 24, pp. 513-523.

SALTON, G. & MCGILL, M. J., 1983. *Introduction to modern information retrieval.* s.l.:McGraw-Hill.

SALTON, G. & YANG, C.-S., 1973. On the specification of term values in automatic indexing. *Journal of documentation,* Volume 29, pp. 351-372.

SAMPIERI, R. H., COLLADO, C. F. & LUCIO, M. P. B., 2013. *Metodologia de Pesquisa.* 5 ed. Porto Alegre: Penso.

SANTORINI, B., 1990. Part-of-speech tagging guidelines for the Penn Treebank Project.

SCHMITZ, M., BART, R., SODERLAND, S. & ETZIONI, O., 2012. *Open language learning for information extraction.* s.l., Association for Computational Linguistics, pp. 523-534.

SELLTIZ, C., WRIGHTSMAN, L. S. & COOK, S. W., 1967. Métodos de Pesquisa nas Relações Sociais..

SHATAKIRTI, M. T., 2011. *Hyperlink based search algorithms-PageRank and HITS,* s.l.: s.n.

SIDDHARTHAN, A., NENKOVA, A. & MCKEOWN, K., 2011. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics,* Volume 37, pp. 811-842.

SIOUTOS, N. et al., 2007. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics,* pp. 30-43.

SIZOV, G., 2010. *Extraction-Based Automatic Summarization: Theoretical and Empirical Investigation of Summarization Techniques.* s.l.:s.n.

SPARCK JONES, K., 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation,* Volume 28, pp. 11-21.

STROBLE, J. K., STONE, R. B., MCADAMS, D. A. & WATKINS, S. E., 2009. *An engineering-to-biology thesaurus to promote better collaboration, creativity and discovery.* s.l., Cranfield University Press.

STRZALKOWSKI, T., 1999. Natural language information retrieval. *Springer Science & Business Media.*

SUMATHY, K. L. & CHIDAMBARAM, M., 2013. Text Mining: Concepts, Applications, Tools and Issues-An Overview. *International Journal of Computer Applications,* Volume 80.

TAVARES, R., 2007. Construindo mapas conceituais. *Ciências & Cognição,* Volume 12, pp. 72-85.

THAKKAR, K. S., DHARASKAR, R. V. & CHANDAK, M. B., 2010. *Graph-based algorithms for text summarization.* s.l., IEEE, pp. 516-519.

TOFFLER, A., 1970. *Future shock.* New York: Amereon Ltd.

TORRES-MORENO, J.-M., 2014. *Automatic text summarization.* s.l.:John Wiley & Sons.

TSENG, S. et al., 2007. A new approach for constructing the concept map. *Computers & Education,* pp. 691-707.

VALERIO, A., LEAKE, D. B. & CAÑAS, A. J., 2008. *Associating documents to concept maps in context.* s.l., s.n.

VASSOLER, G. A., PERIN, W. A. & CURY, D., 2014. *MergeMaps–A computacional tool for merging of concept maps.* s.l., s.n.

VEKIRI, I., 2002. What is the value of graphical displays in learning?. *Educational Psychology Review,* Volume 14, pp. 261-312.

VIDHYA, K. A. & AGHILA, G., 2010. ext mining process, techniques and tools: an overview.. *International Journal of Information Technology and Knowledge Management,* Volume 2, pp. 613-622.

VILLALÓN, J. J. & CALVO, R. A., 2011. Concept Maps as Cognitive Visualizations of Writing Assignments. *Educational Technology & Society,* Volume 14, pp. 16-27.

VUKOTIC, A. et al., 2015. *Neo4j in Action.* s.l.:Manning.

WANG, S. & LIU, L., 2016. *Prerequisite concept maps extraction for automaticassessment.* s.l., International World Wide Web Conferences Steering Committee, pp. 519-521.

WANG, W. M., CHEUNG, C. F., LEE, W. B. & KWOK, S. K., 2008. Mining knowledge from natural language texts using fuzzy associated concept mapping. *Information Processing & Management,* Volume 44, pp. 1707-1719.

WILLETT, P., BARNARD, J. M. & DOWNS, G. M., 1998. Chemical similarity searching. *Journal of chemical information and computer sciences,* Volume 38, pp. 983-996.

WITTEN, I. H. & FRANK, E., 2005. *Data Mining: Practical machine learning tools and techniques.* s.l.:Morgan Kaufmann.

WU, F. & WELD, D. S., 2010. *Open information extraction using Wikipedia.* s.l., Association for Computational Linguistics, pp. 118-127.

XAVIER, C. C., DE LIMA, V. L. S. & SOUZA, M., 2013. *Open Information Extraction based on lexical-syntactic patterns.* s.l., IEEE, pp. 189-194.

YI, N. & LI, H., 2014. *A practical approach for automatically constructing concept map in E-learning environments.* s.l., IEEE, pp. 582-586.

ZOUAQ, A. & NKAMBOU, R., 2009. Evaluating the generation of domain ontologies in the knowledge puzzle project.. *IEEE Transactions on Knowledge and Data Engineering,* Volume 21, pp. 1559-1572.

ZOUAQ, A., NKAMBOU, R. & FRASSON, C., 2007. *Document Semantic Annotation for Intelligent Tutoring Systems: A Concept Mapping Approach.* s.l., s.n., pp. 380-386.

ZUBRINIC, K., KALPIC, D. & MILICEVIC, M., 2012. The automatic creation of concept maps from documents written using morphologically rich languages. *Expert systems with applications,* Volume 39, pp. 12709-12718.

Appendix A
Research on the Manual Construction of Concept Maps

This appendix presents the questionnaire used for collecting data on the manual construction of concept maps.

Identification

Nome: ___________________ Graduate: _______________

Domain on the subject addressed in the text (0 to 100%): _______ %

1. What is the time taken to carry out the activity?

____ h and ____ min.

2. Which is the element most difficult to identify:

□Concept □Relation

3. Were you able to represent all the concepts you wanted on the map?

□Yes □No

4. Were you able to represent all the relations you wanted on the map?

□Yes □No

5. How easy is it to represent a text written by another person on a concept map?

□1 (*easy*) □2 □3 □4 □5 □6 □7 □8 □9 □10 (*difficult*)

6. What is the cognitive effort to accomplish the task?

□Low □Regular □High

7. What is the Language skill level required to perform the task?

□Basic □Intermediate □Advanced

8. What are the sensations observed when performing the task (more than one)?

□Motivating □Stimulating □Relaxing □Demotivating □Tiring □Stressful

9. Does this activity aid learning about the text? Why?

10. If you had access to a tool that automatically performed this activity, would it be useful? Would you lose some benefit that was gained by doing the activity manually?

Appendix B
Research on the Influence that Concept Map has for the Understanding of the Subject addressed in a Text (Quiz A)

This appendix presents the **Quiz A** used for collecting data of the study about the influence that the concept map automatically summarized from a text has on the understanding of the subject addressed in that text.

Quiz A

This research aims to evaluate the information that people can assimilate and extract from a text and a concept map. It is divided in two steps:

- Step 1: *Text* with 5 multiple choice questions for text interpretation;
- Step 2: *Map* with a discursive question for map interpretation.

Identification

Nome: ___________________

Step 1 – Text Comprehension

Answer the questions 1 to 5 according to information presented in the text below:

Biodiversity is the sum of all species on the planet. Some of these species contain important substances that treat several diseases. The most relevant thing about biodiversity is that the rich North needs biodiversity and the poor South has biodiversity. One of the ways to promote a sustainable development is to pay the poor nations to save the forests that they still have. Resources can be extracted but not exhausted. Thus, the environment can be preserved. The Earth belongs to all mankind. Everybody needs to help in the protection of the planet. And there is much to do. We have to fight pollution in all its forms to avoid acid rain, the greenhouse effect, and the death of species, rivers, lakes and seas.

1. According to the text, biodiversity is:

(A) mixing of species which treat different diseases.

(B) the sum of all the planets.

(C) the combination of all substances of species.

(D) the set of all animal and plant species.

2. The biological variety:

(A) exists in the north.

(B) lack in the north.

(C) lack in the south.

(D) enriches the north.

3. The environment can be preserved:

(A) by the increase of forests in rich countries.

(B) with financial assistance to save forests in poor countries.

(C) by the development of poor nations.

(D) by the interruption of resource extraction in poor nations.

4. The protection of the planet depends on:

(A) of all of us.

(B) preservation of forests.

(C) poor countries.

(D) commitment of major industries.

5. One consequence of pollution not mentioned in the text is:

(A) climate change.

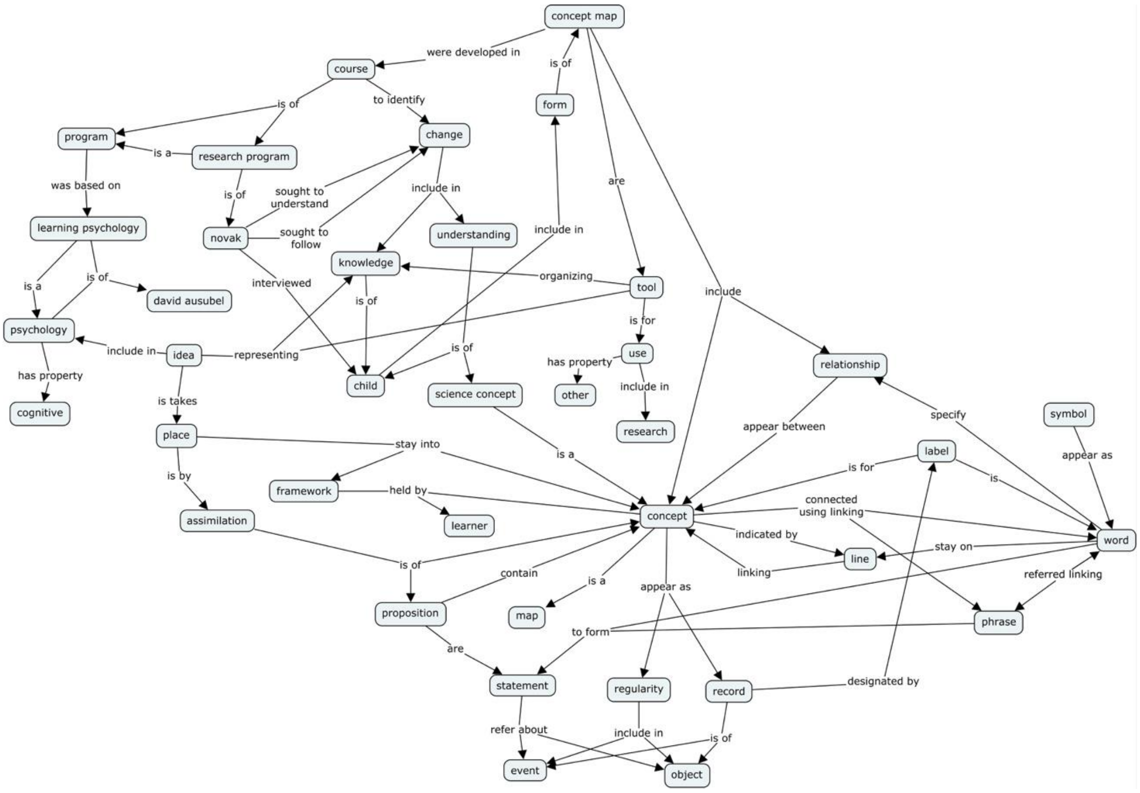
(B) acid rain.

(C) greenhouse effect.

(D) death of species.

Step 2 – Concept Map Comprehension

Make a short summary (5 lines) representing the information presented in the concept map below.


concept map
were developed in
course
is of
to identify
form
program
is a
research program
change
was based on
is of
sought to understand
include in
are
learning psychology
novak
sought to follow
understanding
include in
knowledge
interviewed
organizing
is a
is of
david ausubel
tool
include
is of
psychology
include in
idea
representing
is of
is for
child
use
has property
has property
science concept
other
relationship
include in
cognitive
is takes
research
appear between
specify
symbol
place
stay into
label
appear as
is a
is for
is by
framework
held by
is
learner
connected using linking
concept
assimilation
indicated by
word
stay on
is of
line
contain
linking
is a
referred linking
proposition
appear as
map
to form
phrase
are
statement
regularity
record
designated by
refer about
include in
is of
event
object

Appendix C

Research on the Influence that Concept Map has for the Understanding of the Subject addressed in a Text (Quiz B)

This appendix presents the **Quiz B** used for collecting data of the study about the influence that the concept map automatically summarized from a text has on the understanding of the subject addressed in that text.

Quiz B

This research aims to evaluate the information that people can assimilate and extract from a text and a concept map. It is divided in two steps:

(i) Step 1: *Map* with 5 multiple choice questions for map interpretation;

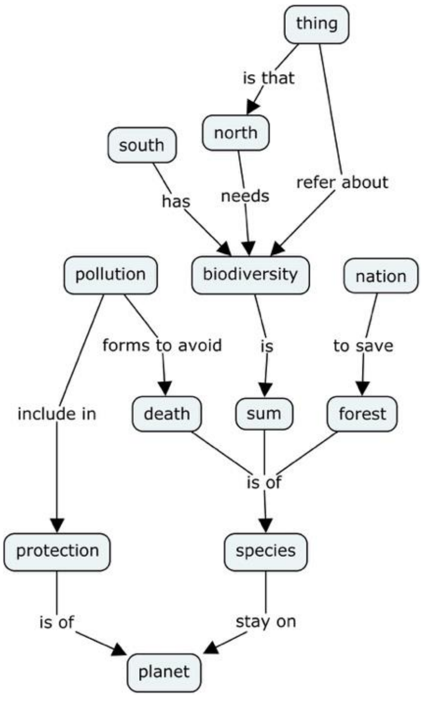
(ii) Step 2: *Text* with a discursive question for text interpretation.

Identification

Nome: ___________________

Step 1 – Map Comprehension

Answer the questions 1 to 5 according to information presented in the map below:



1. According to the text, biodiversity is:

(A) mixing of species which treat different diseases.

(B) the sum of all the planets.

(C) the combination of all substances of species.

(D) the set of all animal and plant species.

2. The biological variety:

(A) exists in the north.

(B) lack in the north.

(C) lack in the south.

(D) enriches the north.

3. The environment can be preserved:

(A) by the increase of forests in rich countries.

(B) with financial assistance to save forests in poor countries.

(C) by the development of poor nations.

(D) by the interruption of resource extraction in poor nations.

4. The protection of the planet depends on:

(A) of all of us.

(B) preservation of forests.

(C) poor countries.

(D) commitment of major industries.

5. One consequence of pollution not mentioned in the text is:

(A) climate change.

(B) acid rain.

(C) greenhouse effect.

(D) death of species.

Step 2 – Text Comprehension

Make a short summary (5 lines) representing the information presented in the text below.

Concept maps are graphical tools for organizing and representing knowledge. Concept maps include concepts, usually enclosed in circles or boxes of some type, and relationships between concepts indicated by a connecting line linking two concepts. Words on the line, referred to as linking words or linking phrases, specify the relationship between the two concepts. We define concept as a perceived regularity in events or objects, or records of events or objects, designated by a label. The label for most concepts is a word, although sometimes we use symbols such as + or %, and sometimes more than one word is used. In the concept map, 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.

Concept maps were developed in 1972 in the course of Novak's research program at Cornell where he sought to follow and understand changes in children's knowledge of science. During the course of this program the Novak interviewed many children, and he found it difficult to identify specific changes in the children's understanding of science concepts by examination of interview transcripts. This program was based on the learning psychology of David Ausubel. The fundamental idea in Ausubel's cognitive psychology is that learning takes place by the assimilation of new concepts and propositions into existing concept and propositional frameworks held by the learner. This knowledge structure as held by a learner is also referred to as the individual's cognitive structure. Out of the necessity to find a better way to represent children's conceptual understanding emerged the idea of representing children's knowledge in the form of a concept map. Thus was born a new tool not only for use in research, but also for many other uses.

1. Short summary (5 lines):