Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 17/Mar/2018; 1st round notif.: 11/Jul/2018; New version: 30/Aug/2018; 2nd round notif.: 08/Oct/2018
Camera ready: 15/Oct/2018; Edition review: 15/Nov/2018; Available online: 30/Nov/2018; Published: 01/Jan/2019

Minerando Mapa Conceitual a partir de Texto em Português

Title: Mining Concept Map from Text in Portuguese

Camila Zacché de Aguiar
Universidade Federal do Espírito Santo, Brasil
camila.zacche.aguiar@gmail.com

Davidson Cury
Universidade Federal do Espírito Santo, Brasil
dedecury@gmail.com

Amal Zouaq
Universidade de Ottawa, Canadá
azouaq@uottawa.ca

Resumo

Mapas conceituais são ferramentas gráficas para representação e construção do conhecimento. A construção manual de um mapa conceitual requer tempo e esforço cognitivo, sendo este acrescido quando o mapa não representa a estrutura cognitiva do autor, mas sim, as informações expressas em um texto escrito por outro autor. Portanto, propomos uma abordagem computacional para mineração de mapas conceituais a partir de textos em português que objetiva representar a informação do texto de forma sumarizada, ou seja, por meio de conceitos e relações aderentes à informação. Para esse fim, definimos arquiteturas, conceitual e tecnológica, que compreendem os serviços de: (i) formatação do texto, removendo caracteres e design do texto; (ii) identificação do domínio, baseado em técnicas de recuperação de informação para identificar o domínio ao qual o texto se refere; (iii) extrator de elementos, usando técnicas de processamento de linguagem natural sobre o texto visando a extração de proposições do tipo conceito-relação-conceito; (iv) sumarizador de elementos, apoiado em análise de grafo para identificar os conceitos relevantes do mapa; e (v) visualização do mapa, por meio da apresentação das proposições de forma gráfica. A abordagem desenvolvida apresenta bons resultados e contribui excepcionalmente para a sumarização de textos, procurando identificar os conceitos relevantes do texto e mantendo suas diversas e mais importantes características. Ademais, esta pesquisa introduz a especificação de um toolkit a fim de prover recursos computacionais para processamento, manipulação e extração de mapas conceituais.
Palavras-Chave: *mapa conceitual, sumarização automática, mineração de mapas conceituais.*

Abstract

Concept maps are graphical tools for representation and construction of knowledge. The manual construction of a concept map requires time and cognitive effort, this being increased when the map should not represent the cognitive structure of the author, but rather, the information expressed in a text written by another author. Therefore, we propose a computational approach for concept map mining from texts in Portuguese that aims to represent the text in summary form through concepts and relationships. To this end, we define a technological architecture that includes the services of: (i) text formatting, removing characters and designing of the text; (ii) domain identification, information retrieval techniques to identify the domain to which refers the text; (iii) elements extractor, natural language processing techniques on the text to extract concept-relation-concept propositions; (iv) element summarizer, supported by graph analysis to identify the relevant concepts on the map; and (v) map visualization, presentation of the propositions in graphic form. The approach developed presents satisfactory results and contributes exceptionally to the summarization of texts to identify the relevant concepts of the text while maintaining its several and most important characteristics. Furthermore, this research introduces the specification of a project to provide computational resources for processing, handling and extraction of conceptual maps.
Keywords: *concept map, automatic summarization, concept map mining*

Cite as: Aguiar, C. Z., Cury, D. & Zouaq, A. (2019). Mining Concept Map from Text in Portuguese (Minerando Mapa Conceitual a partir de Texto em Português). Brazilian Journal of Computers in Education (Revista Brasileira de Informática na Educação - RBIE), 27(1), 83-103. DOI: 10.5753/RBIE.2019.27.01.83

1 Introdução

A sociedade da informação está constantemente acessando informações de forma muito rápida e ampla e novas informações são produzidas, refletidas, publicadas ou compartilhadas quase que instantaneamente. Embora nos permita aprofundar no espaço cognitivo, esta vasta rede de informações produz uma sobrecarga cognitiva reconhecidamente prejudicial aos nossos processos perceptivos e cognitivos, uma indicação de que esses processos são sobrecarregados pelos avanços tecnológicos (Toffler, 1970). Em outras palavras, somos incapazes de absorver e processar toda a informação a que estamos expostos. Analogamente, os dados acadêmicos seguem essa mesma tendência. A partir de uma análise quantitativa sobre as coleções de algumas bases de dados científicas (Springer, IEEE Xplore e ACM), verificamos que o número de artigos publicados nos últimos dezesseis anos é superior ao do último século!

Assim, um estudante é confrontado constantemente com grande quantidade de informação, em fluxo incontrolável, para se manter informado sobre um dado assunto. Esse fato traz diferentes desafios para o processo de aprendizagem do estudante, dos quais destacamos: (i) O estudante deve selecionar os documentos relevantes para um determinado assunto dentre todos aqueles disponíveis. Naturalmente, ele precisa compreender as informações contidas nesses documentos para depois selecioná-los ou não; (ii) Os documentos são compostos por grande quantidade de informação, geralmente escrita de forma complexa e, na maioria dos casos, em idioma diferente do idioma nativo do estudante, o que interfere na sua capacidade de ler e entender; (iii) Depois de selecionar o documento relevante, o estudante também deve investir um grande esforço cognitivo para identificar e entender as informações descobertas.

A pesquisa da literatura sugere que a representação gráfica pode reduzir os problemas de sobrecarga de informação e desorientação de aprendizagem para estudantes (Chen, Wei, & Chen, 2008). Assim, mapas conceituais, enquanto uma representação gráfica, podem ser utilizados como uma representação mais significativa da informação. Portanto, 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 minimizar tempo e esforço cognitivo.

Mapas conceituais utilizados como ferramenta para a representação gráfica de textos proporciona uma maneira visual e holística de representação do conhecimento. Usando mapas conceituais, o estudante poderia conhecer os principais conceitos do assunto antes de se aprofundar no texto. Isso favoreceria a assimilação de novos conhecimentos, especialmente em textos cujo idioma não é o idioma nativo do estudante. Consequentemente, olhando para a representação gráfica, o estudante poderia reduzir o tempo necessário para analisar a relevância do documento para o assunto. Assim, um mapa conceitual que represente efetivamente a sumarização de um documento, permite aos usuários obter uma certa compreensão independente do documento (Karannagoda, et al., 2013).

De acordo com Novak & Cañas (2010), os mapas conceituais são uma ferramenta gráfica para representar e organizar conhecimento compreendendo de conceitos e relações entre eles. Entretanto, a construção manual exige grande dedicação de tempo e esforço mental e físico. Além disso, sua construção torna-se mais complexa quando o autor não representa seu conhecimento, mas o conhecimento expresso em um texto escrito por outra pessoa.

Esta pesquisa tem como objetivo desenvolver uma arquitetura computacional para automaticamente construir mapas conceituais de estilo científico como sumarização de textos acadêmicos. A sumarização proposta não se designa a uma coleção de conceitos, mas sim, a

identificação de conceitos suficientes para representar de forma resumida o texto, mantendo suas diversas e mais importantes características. A arquitetura é suportada por técnicas distintas e complementares em Processamento de Linguagem Natural e Recuperação de Informação. Secundariamente, esta pesquisa direciona esforços no uso de mapas conceituais sumarizados automaticamente como apoio ao processo de aprendizagem dos textos que lhes deu origem.

Este artigo é uma versão estendida do overview apresentado em Aguiar, Cury & Zouaq (2017) cujo resultado se baseia na categorização e revisão da literatura discutidos em Aguiar, Cury & Zouaq (2018), no processo de mineração de mapas conceituais definido em Aguiar, Cury & Zouaq (2016) e na abordagem para extração de elementos a partir de textos apresentado em Aguiar & Cury (2017). Além desta Introdução, a Seção 2 do artigo detalha o método de pesquisa adotado para o seu desenvolvimento; a Seção 3 explora a importância dos mapas conceituais e a aprendizagem definindo dois estilos distintos de mapas conceituais; a Seção 4 discute a representação da informação textual em mapas conceituais; a Seção 5 revisita e propõe alterações ao processo de mineração de mapas conceituais proposto em Aguiar, Cury & Zouaq (2016); a Seção 6 estende a abordagem proposta em Aguiar, Cury & Zouaq (2017) e detalha sua arquitetura tecnológica, bem como apresenta a API produzida como artefato da pesquisa; a Seção 7 apresenta os resultados da pesquisa de construção manual de mapas conceituais; a Seção 8 rediscute a análise da abordagem para a sumarização de textos apresentada em Aguiar, & Cury (2017); a Seção 9 revisita os trabalhos relacionados identificados em Aguiar, Cury & Zouaq (2018) e a Seção 10 apresenta as considerações finais e trabalhos futuros.

2 Método de Pesquisa

Esta pesquisa se iniciou por meio de uma **questão de investigação** assim descrita: "*Como abordagens tecnológicas podem automaticamente construir mapas conceituais sumarizados a partir de textos?*". Dessa questão, definimos um **processo de mineração de mapas conceituais** cobrindo quatro eixos de interesse: Descrição da Fonte de Dados, Definição de Domínio, Extração de Elementos e Visualização do Mapa, detalhada em Aguiar, Cury & Zouaq (2016).

Devido à necessidade de melhor identificar e analisar as funcionalidades e características inerentes à questão de investigação, realizamos um **estudo detalhado** sobre abordagens tecnológicas para construção automática de mapas conceituais publicados entre 1994 e 2016 nas bases de dados IEEE Xplore, ACM e Elsevier Science Direct. A partir desse estudo, elaboramos uma **categorização** definida em duas perspectivas, Fonte de Dados e Representação Gráfica, contendo 14 categorias, detalhada em (Aguiar, Cury, & Zouaq, 2018). O estudo selecionou 30 artigos relevantes, que foram aplicados à categorização proposta para identificar os principais aspectos e limitações de cada abordagem. Desses, apenas cinco foram identificados como **trabalhos relacionados**.

A partir das informações coletadas, elaboramos um **modelo conceitual** para construção automática de mapas conceituais a partir de textos em português, do qual resultou o desenvolvimento de uma **arquitetura orientada a serviços**. Tal arquitetura foi implementada como a ferramenta pública on-line CMBuilder[1], elaborada a partir do Toolkit ExtroutMap[2] que foi produzida como um artefato da pesquisa e disponibilizada publicamente para utilização, extensão e requisição de serviços.

[1] http://cmpaas.inf.ufes.br/cmbuilder
[2] http://extroutmap.inf.ufes.br

Finalmente, para analisar o contexto e os artefatos produzidos, realizamos uma série de **experimentos** e **provas de conceito**: a análise sobre os desafios da construção manual de mapas a partir de textos, a análise comparativa entre o mapa conceitual construído automaticamente pelo CMBuilder e os mapas conceituais construídos manualmente por especialistas e a análise dos especialistas sobre o mapa construído pela abordagem proposta em relação aos mapas construídos pelos trabalhos relacionados. Vale ressaltar que, embora existam trabalhos relacionados, o CMBuilder incorpora características diferenciadas dos trabalhos relacionados, tal como, processamento de artigos inteiros e construção de mapa conceitual como sumarização do texto.

3 Mapas Conceituais e a Aprendizagem

Mapas conceituais são ferramentas gráficas para a representação e construção do conhecimento (Novak & Cañas, 2010), dado que a estrutura cognitiva de um indivíduo pode ser interpretada como uma coleção de conceitos relacionados entre si para formar proposições significativas. Um conceito é definido como uma regularidade percebida em eventos ou objetos, ou registros de eventos ou objetos, designado por um rótulo. Uma proposição é definida como uma declaração significativa sobre um evento ou objeto. Assim, as proposições são formadas a partir da tripla conceito-relação-conceito, constituindo uma unidade semântica.

A Figura 1 apresenta os elementos básicos constitutivos de um mapa conceitual onde os conceitos são representados por retângulos ou caixas e as relações representadas por uma seta rotulada, com direção. Observamos na figura que a organização hierárquica dos conceitos é estabelecida pela posição dos elementos no mapa. Normalmente, os conceitos mais genéricos aparecem no topo do mapa, enquanto os mais específicos aparecem na parte inferior. Além disso, as setas podem indicar a sequência e a direção de como o conhecimento é construído.

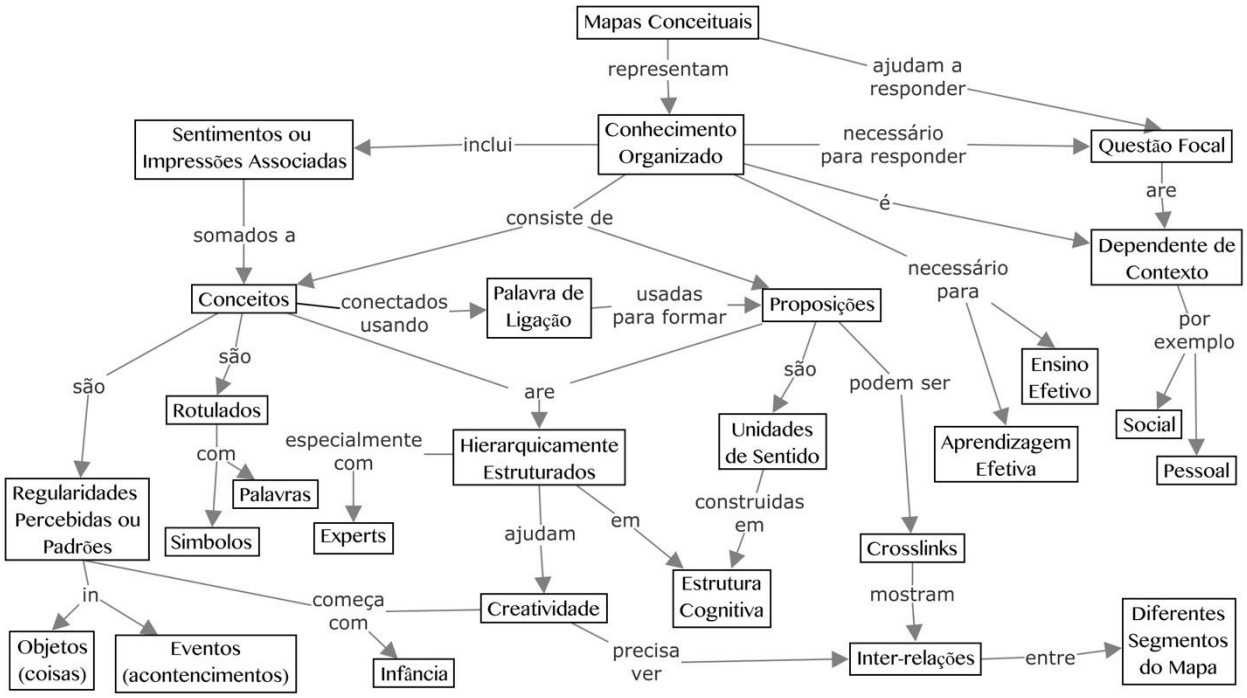

Figura 1: Exemplo de mapa conceitual (Novak & Cañas, 2010).

A partir de nossa experiência pedagógica com mapas conceituais, identificamos e definimos dois estilos distintos de mapas conceituais: Mapa Conceitual de Estilo Científico, foco dessa pesquisa, e de Estilo Educacional.

Definimos por **estilo científico** aquele mapa conceitual construído a partir de uma fonte de dados resultante de uma pesquisa científica, apoiado por duas regras básicas: composto apenas por conceitos e existência de verbo em uma relação entre conceitos. Nesse caso, cada rótulo de conceito consiste em uma ou mais palavras contendo um ou mais substantivos; cada rótulo de relação consiste em uma ou mais palavras contendo um verbo; e cada proposição representa uma unidade de significado. Um mapa conceitual de estilo científico está direcionado a um propósito específico, tal como, avaliação e suporte à aprendizagem, representação e sumarização de textos, entre outros.

Por outro lado, definimos o mapa conceitual de **estilo educacional** como aquele que se destina apenas em representar o conhecimento do indivíduo sem se ater às regras mencionadas, tal como, mapas construídos por crianças para representar seu conhecimento sobre o mundo. Como exemplo, uma criança escreve a frase "*Maria é bonita*". A frase pode ser representada por um simples mapa conceitual contendo a tripla "*Maria-é-bonita"*. No entanto, sabemos que nem "*Maria*" nem "*bonita*" são conceitos. "*Maria*" pode ser definida como uma instância de pessoa ou mulher e, "*bonita*", como uma característica de Maria. No entanto, a frase representa o conhecimento construído por uma criança e é importante ser representado em um mapa conceitual do estilo que definimos como educacional. Entretanto, a frase "*Os professores ensinam certos assuntos*" pode ser representada por um mapa conceitual contendo a tripla "*professores-ensinam-certos assuntos"* que expressa mais claramente a relação significativa entre dois conceitos. Nesse caso, o mapa se caracteriza como do estilo científico.

Podemos considerar vários contextos em que mapas conceituais podem servir como uma ferramenta útil para qualquer teoria da aprendizagem, sendo utilizados como recurso de aprendizagem, meio de avaliação, organização instrucional, representação cognitiva, elicitação ou compartilhamento do conhecimento. Ademais, podem representar com vantagem a informação contida em documentos extensos, uma vez que sua representação gráfica, dinâmica e flexível, em forma de conceitos e relações, é considerada mais fácil de ser construída, assimilada e compreendida do que um texto extenso e gramaticalmente regrado.

4 Representação da Informação Textual em Mapas Conceituais

Informação é conhecimento registrado em forma oral, audiovisual ou escrita, que envolve um elemento de significado (Le Coadic, 1996). Portanto, a informação deve transferir o conhecimento de maneira ordenada e adequadamente estruturada. Caso contrário, permanece inutilizável e amorfo (Mcgarry & De Lemos, 1999). Nesse sentido, a informação explícita promove a assimilação e a interpretação, gerando conhecimento tácito.

Um dos meios mais utilizados para comunicar informações é a linguagem falada ou escrita. Representar as informações corretamente na linguagem escrita é uma tarefa árdua e custosa. Por exemplo, um estudante interessado em representar o conhecimento tácito em uma forma sumária precisaria exercer um grande esforço cognitivo para preparar a síntese. Além disso, a representação exigiria uma organização sequencial, adoção de um estilo, cumprimento das regras gramaticais, preocupação com o formato, entre outros (Gava, Menezes, & Cury, 2003).

A seguir, exemplificamos a diferença de representar informações como um texto escrito e como um mapa conceitual (Figura 2). No texto, a informação que designa um item-chave é representada como um conceito no mapa, dentro de uma caixa. As informações que indicam uma ação ou evento são representadas como uma relação no mapa, como uma seta direcional rotulada. Além disso, notamos que o mapa conceitual não representa todas as informações do texto, mas apenas aquelas que formam proposições significativas.

"Mapas conceituais são ferramentas gráficas para organizar e representar o conhecimento. Eles incluem conceitos, geralmente colocados em círculos ou caixas de algum tipo, e relacionamentos entre conceitos indicados por uma linha de conexão que relaciona dois conceitos".

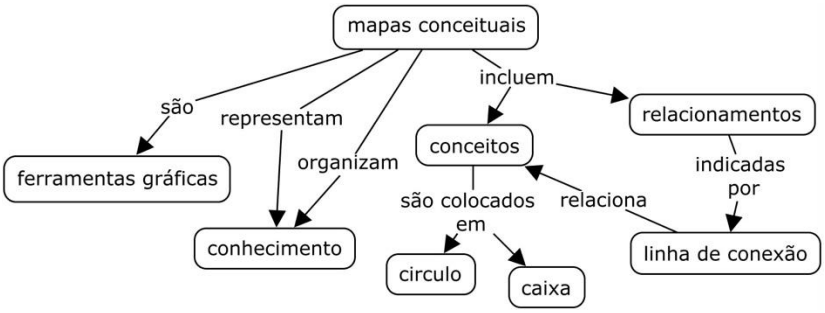

Figura 2: Texto escrito extraído a partir de (Novak & Cañas, 2010) e sua representação como mapa conceitual

Um texto escrito aderente às regras gramaticais pode ser representado por um mapa conceitual de forma gráfica e holística. Em outras palavras, uma representação gráfica mais dinâmica e flexível do texto pode ser construída, o que facilita a assimilação e compreensão do texto original. Além disso, proposições significativas permitem ao leitor obter um novo ponto de vista sobre as informações essenciais expressas no texto. Assim, um único mapa pode ser interpretado de diferentes maneiras, dependendo do leitor, assim como, um único texto pode gerar diferentes mapas, dependendo do autor.

5 Processo de Mineração de Mapas Conceituais

A Mineração de Mapas Conceituais (Concept Map Mining - CMM) é definida como um processo para a automática extração de mapas conceituais a partir de documentos em contexto educacional (Villalón & Calvo, 2011). O processo proposto por Aguiar, Cury & Zouaq (2016) engloba quatro eixos de interesse, tal processo é adotado nesta pesquisa acrescido da etapa de Sumarização de Elementos: (i) **Descrição da Fonte de Dados**, que define as técnicas utilizadas no processo de extração de informações, bem como os métodos de manipulação apropriados, tais como técnicas linguísticas, estatísticas, aprendizagem de máquina, recuperação e identificação de elementos; (ii) **Definição do Domínio**, que identifica o domínio do documento e, consequentemente, possibilita identificar seus conceitos relevantes; (iii) **Identificação de Elementos**, que extrai conceitos e relações para a formação de proposições; (iv) **Sumarização de Proposições**, seleciona as proposições relevantes segundo o domínio do texto; e (v) **Visualização do Mapa**, que especifica o posicionamento gráfico das proposições no mapa conceitual, uma vez que os mapas são ferramentas gráficas e a visualização do conhecimento é parte do processo de aprendizagem.

O processo é apresentado na Figura 3 e pode ser formalizado e descrito pelos seus elementos constituintes: Descrição da Fonte de Dados, Definição do Domínio, Identificação de Elementos, Sumarização de Proposições e Visualização do Mapa.

A **Descrição da Fonte de Dados** caracteriza um documento *D*. Um documento *D* de tamanho *n* pode ser definido como $D = \{d_1 ... d_n\}$ onde d_i, $i=1...,n$ é um termo em *D*. A definição de um processo de construção de mapa conceitual está relacionada à fonte de dados que será usada como fonte para extração do mapa. Nesse contexto, identificamos variações nas abordagens no que se refere ao tamanho (pequeno, regular e longo) e quantidade (único e conjunto de documentos) de informações disponíveis na fonte de dados, detalhadas e classificadas em (Aguiar, Cury, & Zouaq, 2018).

O documento D é usado como entrada na etapa de **Definição do Domínio** para a descoberta do domínio Ω do documento. O domínio Ω é a união dos conceitos C relacionados a um mesmo assunto, extraídos a partir de documentos D. Dado o desafio em identificar o domínio do texto e dos conceitos que pertencem ao texto, algumas abordagens fazem uso de técnicas semiautomáticas onde o autor identifica o domínio da fonte de dados selecionando ontologia, mapas, conceitos ou documentos. Essa etapa torna-se essencial para a construção de mapas conceituais que representem o domínio expresso no texto.

Um conjunto de conceitos pode ser definido como $C = \{c_1...c_n\}$ onde $C \subseteq D$ e c_i é um termo d_i que representa um conceito ou entidade para o domínio. Um conjunto de relações pode ser definido como $R = \{r_1...r_n\}$ onde $R \subseteq D$ e r_i é um conjunto de termos concatenados que representa uma relação entre conceitos. Uma proposição pode ser definida como $P_{ijk} = \{c_i, r_j, c_k\}$ onde $c_i \in C$ e $c_k \in C$ e $r_j \in R$. A etapa de **Identificação dos Elementos** objetiva extrair as proposições, ou seja, triplas conceito-relação-conceito, que irão compor o mapa conceitual. No entanto, sua extração automática ainda é um desafio e resulta em mapas fragmentados contendo conceitos desconectados, rótulos incompletos ou extensivos, ou frases de ligação não identificadas.

O conjunto de proposições $P = \{p_1...p_n\}$ é sumarizado σ na etapa de **Sumarização de Proposições** a fim de filtrar aquelas proposições relevantes ao domínio do texto e excluir aquelas não significativas. Esta etapa é de grande importância para a qualidade do mapa conceitual resultado, uma vez que um mapa ilegível ou não representativo contribui pouquíssimo para o aprendizado do leitor.

Durante a etapa de **Visualização do Mapa**, para cada proposição P_{ijk}, atribuímos uma posição gráfica G_i para formar um conjunto de proposições organizadas topologicamente em um mapa conceitual definido como $CM = \{P_{ijk}, G_i\}$. Consideramos que essa etapa se faz tão importante como as outras, uma vez que o aprendizado se dá pela representação da informação por meio de conceitos e relações, e não, como texto.

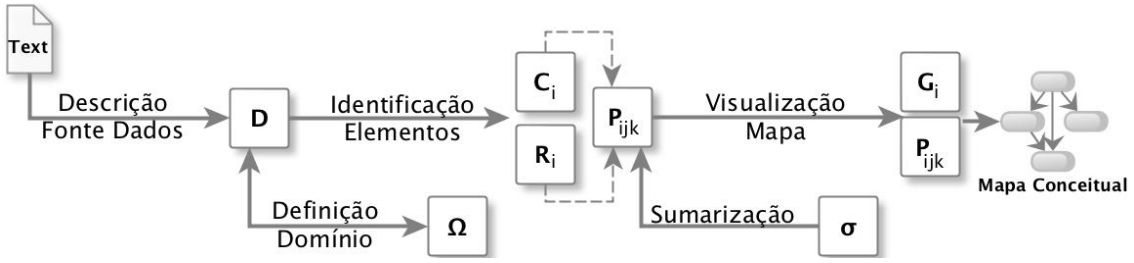

Figura 3: Processo de mineração de mapas conceituais

6 Minerando Mapa Conceitual a partir de Texto

Nesta seção apresentamos uma abordagem para mineração de mapas conceituais a partir de textos apoiada no processo proposto na Seção 5. A Figura 4 apresenta uma visão geral da abordagem e seus componentes, desenvolvida sobre um ambiente web.

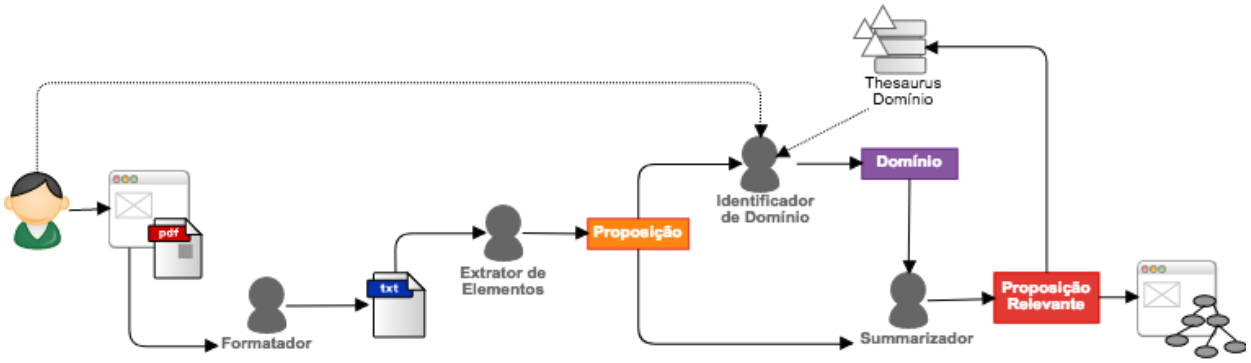

Figura 4: Modelo conceitual da abordagem

Uma síntese do processo mostrado na figura pode ser descrita da seguinte forma:

- O modelo descreve a **Fonte de Dados** como artigos científicos de idioma português, ou seja, a fonte de dados é de tamanho regular formada por algumas páginas, nem de tamanho pequeno formada por poucos parágrafos como resumos e nem de tamanho grande formada por muitas páginas como jornais e teses. O usuário acessa uma aplicação web e carrega uma fonte de dados em formato pdf.
- Em seguida, o **Serviço Formatador** transforma esse pdf em um texto não formatado e pré-processado.
- A partir desse texto, o **Serviço Extrator de Elementos** extrai um conjunto de proposições usando um dicionário léxico e processamento linguístico.
- Em seguida, as proposições são usadas pelo **Serviço Identificador de Domínio** para encontrar uma referência de domínio no Thesaurus de Domínio. O **Thesaurus de Domínio** é responsável por armazenar os dados sobre os domínios dos textos processados, ou seja, o vocabulário conceitual de um domínio definido.
- As proposições e os conceitos do domínio de referência são usados pelo **Serviço Sumarizador** para gerar um mapa conceitual contendo proposições relevantes ao domínio do texto.
- Por fim, tais proposições são devolvidas à aplicação web onde o mapa conceitual é apresentado e editável em uma interface gráfica, cujos elementos são posicionados para **Visualização**.

O modelo conceitual definido foi implementado em uma arquitetura tecnológica contendo os quatro servidores principais da abordagem. Ademais, a arquitetura fez uso dos seguintes recursos tecnológicos: Stanford CoreNLP, Apache OpenNLP e ExtroutMap, descritos a seguir.

Stanford CoreNLP: um pipeline extensível que fornece análise de linguagem natural, tais como Tokenização, Divisão de Sentenças, Tagging de Part-of-speech, Análise Morfológica, Reconhecimento de Entidades Nomeadas, Análise Sintática e Resolução de Correferência (Manning, et al., 2014). O toolkit funciona com modelos em diferentes idiomas, como o inglês. Está disponível como GNU General Public License em http://stanfordnlp.github.io/CoreNLP.

Apache OpenNLP: é um toolkit para processamento de linguagem natural baseado em aprendizado de máquina, capaz de executar Tokenização, Segmentação de Sentenças e Tagging de Part-of-Speech. O toolkit funciona com modelos em diferentes idiomas, como o português. Está disponível como Apache License 2.0 em https://opennlp.apache.org.

ExtroutMap: é uma api desenvolvida durante esta pesquisa que fornece um conjunto de bibliotecas baseadas em técnicas de processamento de linguagem natural e recuperação de informação direcionadas as necessidades de mapas conceituais, tais como Extração de

Informação Aberta, Ranking de Conceitos e Sumarização de Mapas para idiomas em português e inglês. Está disponível como GNU General Public License em http://extroutmap.inf.ufes.br.

A api propõe oferecer um conjunto de bibliotecas em linguagem Java, bem como serviços web para serem requisitados e estendidos por qualquer ferramenta. Os seguintes recursos foram desenvolvidos e estão disponíveis:

- **OpenIE**: Serviço para Extração de Informação Aberta (*Open Information Extration*) em idioma português e inglês. Embora OpenIE possa ser utilizado em vários contextos, introduzimos essa estratégia em extração de triplas a partir de textos para a representação de proposições em mapas conceituais.
- **ConceptRank**: Serviço para ordenar os conceitos pertencentes a um mapa conceitual atribuindo um peso para cada conceito segundo uma métrica.
- **NodeClassify**: Serviço para classificar os *nodes* de um grafo segundo sua estrutura. Embora *NodeClassify* possa ser utilizado em vários contextos, introduzimos essa estratégia para classificar os conceitos relevantes em um mapa conceitual.

Os recursos são projetados em arquitetura orientada a serviços provendo modularização, facilidade de acesso pela internet, extensibilidade e, principalmente, integração. A integração é uma característica fundamental, uma vez que o projeto é uma iniciativa de colaboração e será formado por diversos serviços computacionais frutos de pesquisas distintas relacionadas a mapas conceituais. O modelo orientado a serviços adotado na api é arquitetado em três camadas como segue:

- **Camada de Apresentação,** é a interface de comunicação entre o serviço e a aplicação cliente a fim de requisitar a execução de um serviço e retornar seu resultado. Qualquer aplicação web pode realizar requisições aos serviços ExtroutMap usando protocolo HTTP. Nesta pesquisa o CMBuilder é uma aplicação web que objetiva construir mapas conceituais a partir de textos utilizando os serviços oferecidos pela api ExtroutMap.
- **Camada de Serviço,** é responsável por publicar serviços e realizar a comunicação com a camada de dados. Os serviços são gerenciados e disponibilizados para acesso e extensão, ou seja, outros projetos e ferramentas podem usufruir das pesquisas realizadas e contribuir com melhorias ou especializações;
- **Camada de Dados**, armazena os dados utilizados pelos serviços uma vez que alguns dos serviços publicados pela api ExtroutMap necessitam armazenar informações temporárias ou permanentes.

De forma geral, uma Aplicação Cliente na **Camada de Apresentação** realiza uma requisição sobre um dos serviços disponibilizados pelo ExtroutMap na **Camada de Serviço**. Esse, quando necessário, acessa a **Camada de Dados** e retorna à **Aplicação Cliente** o resultado da requisição. Além disso, outros serviços, em outros servidores, na Camada de Serviços, podem usufruir dos recursos oferecidos realizando requisição a um dos serviços disponibilizados.

As seções subsequentes abordam mais detalhes de cada um dos servidores apresentados no modelo conceitual. Considere como fonte de dados o texto exemplo extraído de Novak & Cañas (2010): "*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 (NOVAK; MUSONDA, 1991). 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 (1963, 1968; AUSUBEL et al.,*

1978). 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.".

6.1 Formatador

O Servidor Formatador é responsável por transformar o arquivo pdf de entrada em um texto não formatado. Para isso, são usadas atividades para extrair texto a partir do pdf; eliminar marcadores de rótulos, referências, tags e estilo da fonte; remover caracteres especiais; e remover frases não-proposicionais, ou seja, aquelas que expressam ordem, questão ou dúvida.

A seguir apresentamos o texto exemplo retornado após ser processado pelo servidor formatador: "*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.".*

6.2 Extrator de Elementos

O Servidor Extrator de Elementos inicia com a etapa de Normalização sobre o texto sem formatação e finaliza com a Extração de Triplas no formato conceito-relação-conceito.

A etapa de **Tokenização**, conversão de uma sequência de caracteres em uma sequência de unidades de significado (palavras), é realizada por meio do módulo *Tokenizer* com o modelo *pt-token* treinado com o *corpus Bosque CoNLL-X*, provido pela ferramenta *Apache OpenNLP*. *Bosque CoNLL-X* é um subconjunto da Floresta Sintática, *treebank* português construído a partir dos *corpora* Jornal Público de Portugal e Folha de São Paulo do Brasil, contendo 9368 sentenças (Afonso, Bick, Haber, & Santos, 2002). A **Análise Morfológica**, identificação da classe gramatical da palavra, é realizada pelo módulo *PosTagger* com modelo *pt-tagger-macmorpho* treinado com o *corpus MacMorpho*, provido pela api ExtroutMap. *MacMorpho* é um *corpus* de textos brasileiros anotado com *tags part-of-speech* (Fonseca & Rosa, 2013). Tanto os *corpora Bosque* como *MacMorpho* definem seu próprio conjunto de *tags*, sendo mapeados para as tags definidas no *CINTIL Treebank*. *CINTIL Treebank* é um corpus português anotado com a representação das relações constituintes (Branco, 2010).

Em seguida, o texto é dividido em sentenças individuais para análise sintática com o auxílio das ferramentas *Apache OpenNLP* e *Stanford CoreNLP*. A **Segmentação do Texto** é realizada por meio do módulo *SentenceDetector* do *OpenNLP* com o modelo *pt-sent* treinado com o *corpus Bosque CoNLL-X*, provido pela ferramenta. A **Análise Sintática**, definição da função que a palavra adota dentro de uma sentença de acordo com certa teoria gramatical, é realizada por meio do módulo *LexicalizedParser* do *CoreNLP* com modelo *pt-parser-cintil* treinado com *corpus CINTIL Treebank*, provido pela api ExtroutMap.

A partir da árvore *parser* produzida pela etapa anterior, são extraídas as triplas com o auxílio do módulo *OpenIE* da api ExtroutMap. O módulo *OpenIE* adota busca em profundidade e regras heurísticas sobre a árvore *parser* para extração das triplas seguindo as etapas de Identificação de Estruturas Independentes, Ajuste de Estruturas e Extração de Triplas.

Cada árvore *parser* é segmentada em um conjunto de **estruturas independentes completas** contendo uma estrutura menos complexa. Definimos por estrutura independente

completa, aquela formada por sintagmas completos seguindo o padrão (1) e (2). Os sintagmas completos são: (i) sintagma completo nominal (NP), composto por um núcleo nominal (NN), ou derivado, (ii) sintagma completo verbal (VP), composto por um núcleo verbal (VB), ou derivado, e um sintagma completo nominal, e (iii) sintagma completo preposicional (PP), composto por um núcleo preposicional (IN), ou derivado, e um sintagma completo nominal. Estruturas intermediárias, sintagmas incompletos e *tags* existentes entre os sintagmas completos, são ignorados.

$$S < ((NP < (NN+)) \$ (VP < (VB+ \$ (NP < (NN+))) \quad (1)$$

$$S < ((NP < (NN+)) \$ (PP < (IN \$ (NP < (NN+))) \quad (2)$$

Essas estruturas independentes são **Ajustadas** onde (i) regras morfológicas são aplicadas para identificar o núcleo dos sintagmas adotando *tokens* nominais e adjetivos para nomes e *tokens* verbais, preposicionais e adverbiais para relações; (ii) todos os tokens que pertencem ao núcleo nominal são lematizados; (iii) cada sintagma preposicional é transformado em um sintagma verbal por meio de um mapeamento; e (iv) relação de especialização entre conceitos é identificada por meio de nomes compostos e estrutura gramatical.

Finalmente, as proposições na forma conceito1-relação-conceito2 são extraídas a fim de representar o fato expresso na estrutura independente completa. Para isso localizamos o primeiro sintagma verbal da estrutura e extraímos: o sujeito (sintagma nominal localizado antes do VP); o objeto (sintagma nominal localizado dentro do VP); e o predicado (*tokens* localizados entre o sujeito e objeto). A partir do sujeito, predicado e objeto é formada a proposição conceito-relação-conceito.

A Tabela 1 apresenta o texto exemplo retornado após ser processado pelo servidor extrator de elementos, ou seja, apresenta o conjunto de triplas extraídas.

Tabela 1: Triplas extraídas pelo Servidor Extrator de Elementos.

Conceito	Relação	Conceito
mapas conceituais	foram desenvolvidos em	1972
programa	compreendiam	ciência
programa	é de	pesquisa
pesquisa	é por	novak
novak	é na	universidade de cornell
novak	entender	mudanças
mudanças	é na	maneira
maneira	aparece como	crianças
pesquisadores	entrevistaram	número
número	é de	crianças
pesquisadores	tiveram	dificuldade
pesquisadores	identificar	mudanças específicas
compreensão	examinando	entrevistas transcritas
compreensão	é de	conceitos científicos
conceitos científicos	é por	parte
parte	é das	crianças
programa	baseava	psicologia
psicologia	é da	aprendizagem
aprendizagem	é de	david ausubel
ideia fundamental	é na	psicologia cognitiva
psicologia cognitiva	é de	david ausubel
aprendizagem	é por meio	assimilação
assimilação	é de	novos conceitos
proposições	permanecer dentro de	conceitos preexistentes

6.3 Identificador de Domínio

O servidor é responsável por identificar o domínio do texto a partir da lista de proposições extraídas na etapa anterior. Um *thesaurus* de domínio é construído automaticamente nesta etapa, armazenando as proposições do texto em grupos de domínio de forma incremental à medida que novos textos do mesmo domínio são processados para criar mapas.

Para identificar o domínio, o servidor realiza cálculo de similaridade de cosseno dos conceitos extraídos com os conceitos pertencentes aos domínios armazenados no Thesaurus. Caso exista algum domínio com similaridade alta (superior a 70%), o mapa é associado para o domínio, senão um novo domínio é criado no Thesaurus. Considerando que o texto exemplo é o primeiro texto processado pela abordagem, o *thesaurus* do domínio se encontra vazio e, portanto, o domínio não é identificado.

6.4 Sumarizador

O servidor aplica uma sumarização sobre as proposições extraídas a fim de selecionar as mais relevantes ao domínio do texto para a construção do mapa conceitual.

A etapa de ***Ranking*** é responsável por atribuir um peso para os conceitos segundo algum parâmetro, tal como, frequência do conceito no texto. Para isso, representamos a lista de conceitos na forma de um grafo, considerando que cada vértice possui um escore *hub*, número de conexões de saída; e escore *authority*, número de conexões de entrada. Assim, o peso *W* de cada conceito *k* é computado pela fórmula (3), cujo peso *W(k)* máximo é igual a 1. A fórmula associa o score *hub H(k)* e *authority A(k)* com a frequência dos conceitos no texto $TF_d(k)$ e no domínio *TFIDF(k).* Os melhores parâmetros de ajuste do modelo HARD (Leake, Maguitman, & Reichherzer, 2004) foram atribuídos para $\rho = 2.235$ and $\sigma = 1.764$. Os parâmetros $\beta = 0.1$, $\alpha = 0.2$, e $\gamma = 0.7$ foram adotados a partir de experimentos não científicos para teste.

$$\boldsymbol{W(k) = [\beta . TFIDF_{\Omega}(k)] + [\alpha . TF_d(k)] + [\gamma . (\rho . A(k) + \sigma . H(k))]} \qquad (1)$$

A etapa de **Sumarização** é responsável por identificar as proposições relevantes diante do conjunto de triplas extraídas. Para isso aplicamos o conceito de quartis à topologia do grafo a fim de classificar os vértices, cujo peso de cada um é atribuído de acordo com a etapa de Ranking. Cada vértice é classificado como *heavy*, caso esteja localizado no terceiro quartil; *interjacent,* caso esteja localizado no caminho entre dois vértices *heavy*; *adjacent*, caso o peso do vértice seja superior ou igual ao menor peso dos vértices *interjacent*; e *light*, caso não se enquadre em nenhuma das classificações anteriores.

A seguir, apresentamos na Figura 5 o texto exemplo retornado após ser processado pelo servidor sumarizador. Dado que o texto exemplo é formado apenas por algumas frases, o servidor sumarizador não possui informação suficiente para identificar com clareza a relevância dos conceitos, resultando em um mapa conceitual simplório.

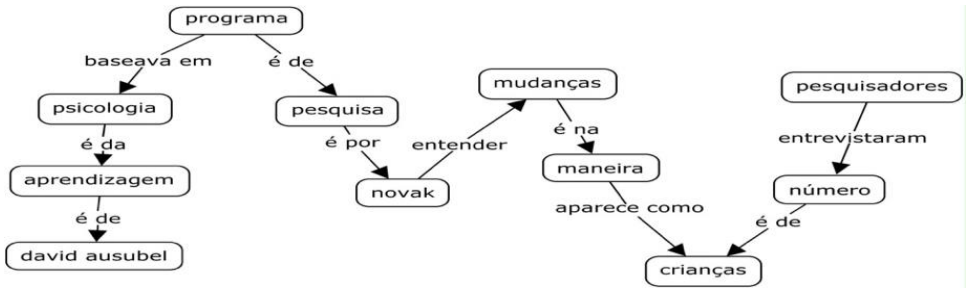

Figura 5: Proposições após o Servidor Sumarizador

7 Desafios da Construção Manual de Mapas Conceituais a partir de Texto

A fim de compreender o processo e as dificuldades da construção manual de mapas conceituais a partir de textos, realizamos uma pesquisa com 10 especialistas no domínio de mapas conceituais. A coleta de dados foi realizada por meio de um questionário[3] contendo 12 perguntas fechadas e abertas que visa identificar as dificuldades no processo de construção manual de mapas conceituais a partir de textos.

A pesquisa foi conduzida da seguinte forma: (i) Os especialistas receberam informações sobre o uso de mapas conceituais e sobre o objetivo da pesquisa; (ii) receberam um texto em inglês contendo 592 palavras, que é o mesmo que foi aplicado no experimento realizado na Seção 7; (iii) foram instruídos a construir um mapa conceitual de natureza essencialmente científica utilizando exclusivamente elementos daquele texto, ou seja, os rótulos dos conceitos devem conter substantivos e os rótulos das relações devem conter verbos; e (iv) após a construção manual do mapa conceitual, os especialistas foram instruídos a responder um questionário.

A partir da pesquisa realizada com os especialistas, coletamos e destacamos algumas informações sintetizadas a seguir. A Figura 6 mostra o tempo gasto pelos especialistas para construir o mapa conceitual a partir do texto. Como mostrado pelo gráfico, o tempo médio gasto para construir manualmente o mapa conceitual foi de 1 hora e 47 minutos.

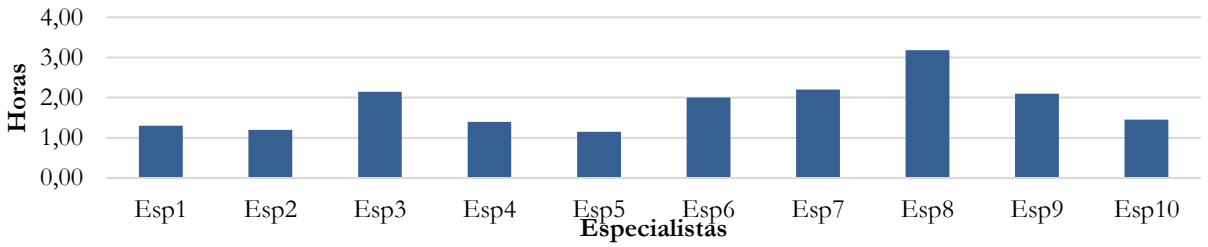

Figura 6: Tempo gasto pelos especialistas

A Figura 7 mostra o nível de facilidade identificado pelo especialista para realizar a tarefa, ou seja, construir o mapa conceitual a partir de um texto. Como mostrado pelo gráfico, a tarefa foi considerada no nível médio de facilidade 4.5.

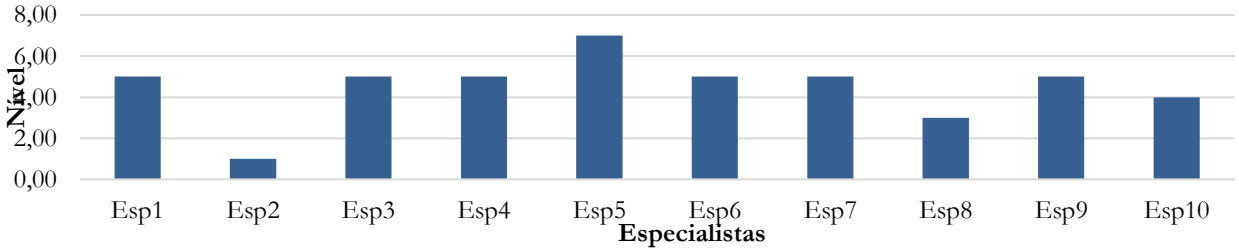

Figura 7: Nível de facilidade para construir o mapa conceitual

Ao analisar os dados, extraímos algumas informações importantes que orientam o desenvolvimento desta pesquisa, como segue:

[3] Questionário e resultados disponíveis em http://extroutmap.inf.ufes.br

- O tempo médio para a construção manual de um mapa conceitual a partir de um texto contendo cerca de 500 palavras é maior que 1 hora.
- A maior dificuldade durante a construção do mapa está relacionada à identificação das relações (100%) e, em alguns casos, o autor não consegue representar todos os conceitos (62,5%) e relacionamentos (75%) que considera relevante.
- Os autores (100%) consideram que esta tarefa requer alto esforço cognitivo e nível intermediário de habilidade com o idioma do texto. A maioria (60%) considera que essa tarefa tem um grau de facilidade 5, em uma escala de 1 a 10.
- Os autores classificaram essa atividade como cansativa (75%), motivadora (75%), estimulante (37,5%) e estressante (25%).
- A tarefa auxilia na compreensão do texto (75%), pois o autor deve (i) ler e reler o texto para extrair conceitos, (ii) aprofundar o entendimento do texto para extrair relações e (iii) encontrar a maneira adequada de representar e conectar os conceitos no mapa.
- Embora alguns autores (25%) considerem que o uso de uma ferramenta tecnológica para construção automática do mapa a partir do texto poderia levar a perdas da informação, todos os autores (100%) concordam que a ferramenta traria grandes benefícios auxiliares para a compreensão do texto.

8 Construção Automática de Mapas Conceituais como Sumarização de Textos

Sumarização é uma representação breve e precisa de um texto de entrada do tipo que a saída cobre os conceitos mais importantes da fonte de dados de forma condensada (Thakkar, Dharaskar, & Chandak, 2010). De acordo com (Hutchins, 1987), os sumários científicos podem ser classificados em três tipos: (i) indicativo, contendo apenas os tópicos essenciais de um texto; (ii) informativo, contendo todos os aspectos principais do texto, considerado como um substituto para o texto; e (iii) avaliativo, apresentando uma análise comparativa entre o conteúdo da fonte de texto e outros trabalhos relacionados.

Esta pesquisa está interessada em sumários científicos de tipo informativo. Portanto, a sumarização automática de texto é uma tarefa que cria uma representação compacta de um documento ou coleção de documentos para entender e cobrir seu objetivo principal. Nesse caso, propomos uma **sumarização automática de texto representada por meio de conceitos e relações na forma de um mapa conceitual**, cobrindo os conceitos relevantes do domínio do texto.

Dada a abordagem proposta (Seção 6), direcionamos esforços para sumarizar um texto científico na forma de mapa conceitual. Não estamos interessados em criar a representação fiel das proposições existentes no texto, e sim uma representação sumária dele.

Para isso, realizamos um experimento utilizando como fonte de dados a seção introdutória do artigo Novak & Cañas (2008), escrito no idioma português e composto por 26 sentenças e 592 palavras, detalhado em (Aguiar & Cury, Mineração de Mapas Conceituais a partir de Textos em Português, 2017). O texto foi processado pela abordagem, identificando 53 conceitos e 95 proposições relacionadas, representados na Figura 8. Os conceitos na cor azul foram considerados irrelevantes pelos especialistas segundo a analise apresentada a seguir.

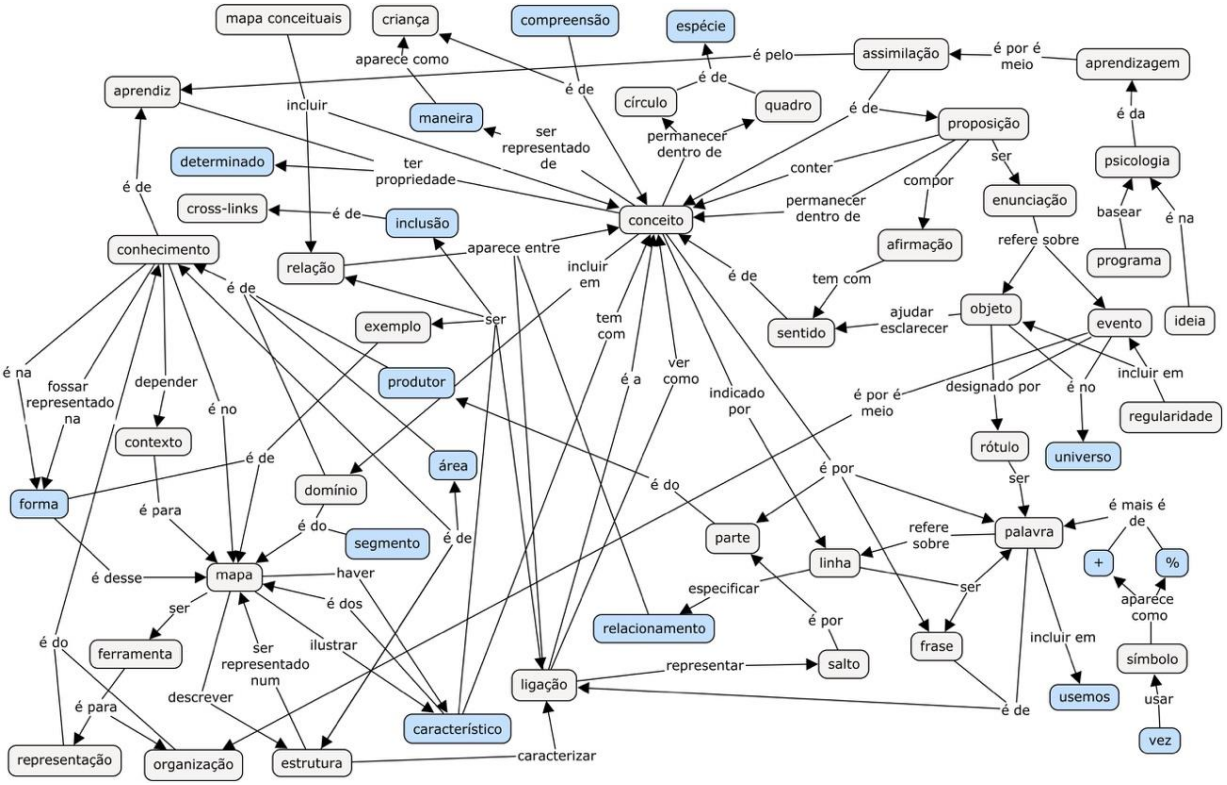

Figura 8: Mapa conceitual sumarizado pela abordagem

Para analisar a fidelidade do mapa conceitual com o texto, comparamos o mapa gerado automaticamente por nossa abordagem com mapas construídos manualmente por especialistas do domínio. Definimos por *especialistas*, estudantes de pós-graduação em informática na educação conhecedores de mapas conceituais. As seguintes instruções foram providas: (i) os especialistas receberam informações sobre o uso de mapas conceituais em geral e sobre o propósito do experimento; (ii) foram instruídos que tanto o rótulo de conceitos quanto de relacionamentos deveria ser curto, significativo e extraído a partir do texto; e (iii) foram informados que rótulos dos conceitos deveriam conter substantivos, e rótulos das relações deveriam conter verbos.

As tabelas a seguir mostram o *precision* e *recall* calculados comparando o mapa construído pela abordagem com os mapas construídos pelos especialistas. A Tabela 2 mostra a análise dos conceitos identificados, alcançando 0.68 em *precision* e 0.38 em *recall*. Neste experimento desconsideramos a flexão do rótulo dos conceitos, tal como plural.

Tabela 2. Score da fidelidade dos Conceitos

Especialista		**Esp.1**	**Esp.2**	**Esp.3**	**Esp.4**	**Esp.5**	**Média**
Abordagem	*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**

A Tabela 3 mostra a análise sobre as relações identificadas, obtendo 0.41 em *precision* e 0.19 em *recall*. Nesta avaliação, consideramos como relações similares aquelas que conectam os mesmos conceitos, exatamente, e que possuem significado aproximado.

Tabela 3. Score da fidelidade das Relações

Especialista		Esp.1	Esp.2	Esp.3	Esp.4	Esp.5	Média
Abordagem	*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**

O valor baixo alcançado pela métrica *recall* pode ser explicado devido o tamanho do mapa conceitual construído pelos especialistas. Uma vez que os especialistas estavam lendo o texto em seu idioma nativo e tinham domínio sobre o assunto, os mapas construídos foram muito breves e com mínimo conjunto de conceitos.

Além deste ponto, podemos destacar alguns outros relevantes que influenciaram o resultado: (i) Alguns rótulos atribuídos pela abordagem não corresponderam aos rótulos atribuídos pelos especialistas. A abordagem, às vezes, não faz uso de alguns adjetivos e advérbios relativamente importantes para caracterizar os conceitos; (ii) Alguns relacionamentos atribuídos pelos especialistas não foram explicitamente extraídos do texto porque a informação pré-existente em sua estrutura cognitiva interferiu na construção do mapa; (iii) Alguns conceitos relevantes foram perdidos durante o segundo experimento devido a etapa de *ranking* e sumarização; e (iv) A atividade de extrair proposições a partir do texto e não a partir do conhecimento prévio do autor, requer muito tempo e grande esforço cognitivo, fato que prolongou a execução da atividade por mais de uma hora, afetando a qualidade dos mapas.

Portanto, para verificar a qualidade dos mapas conceituais construídos pelos especialistas, realizamos uma análise quantitativa comparando o mapa de cada especialista com todos os demais. A Tabela 4 mostra a análise sobre os conceitos identificados, alcançando score médio de 0.63 em *precision* e *recall*, inferior ao score obtido pela abordagem (Tabela 2).

Tabela 4. Score dos Conceitos identificados pelos Especialistas

Especialista		Esp.1	Esp.2	Esp.3	Esp.4	Esp.5	Média
Esp.1	*Precision*	0.00	0.85	0.70	0.71	0.74	**0.75**
	Recall	0.00	0.28	0.35	0.85	0.50	**0.49**
Esp.2	*Precision*	0.28	0.00	0.45	0.27	0.44	**0.36**
	Recall	0.85	0.00	0.69	1.00	0.92	**0.87**
Esp.3	*Precision*	0.35	0.69	0.00	0.38	0.56	**0.49**
	Recall	0.70	0.45	0.00	0.90	0.75	**0.70**
Esp.4	*Precision*	0.85	1.00	0.90	0.00	0.81	**0.89**
	Recall	0.71	0.27	0.38	0.00	0.46	**0.45**
Esp.5	*Precision*	0.50	0.92	0.75	0.46	0.00	**0.66**
	Recall	0.74	0.44	0.56	0.81	0.00	**0.64**

A Tabela 5 mostra a analise sobre as relações identificadas, alcançando score médio de 0.42 em *precision* e *recall*, próximo ao obtido pela abordagem (Tabela 3).

Tabela 5. Score da Relações identificadas pelos Especialistas

Especialista		Esp.1	Esp.2	Esp.3	Esp.4	Esp.5	Média
Esp.1	*Precision*	0.00	0.53	0.63	0.55	0.43	**0.53**
	Recall	0.00	0.14	0.27	0.71	0.23	**0.34**
Esp.2	*Precision*	0.14	0.00	0.33	0.11	0.30	**0.22**
	Recall	0.53	0.00	0.53	0.53	0.60	**0.55**
Esp.3	*Precision*	0.27	0.53	0.00	0.21	0.37	**0.34**
	Recall	0.63	0.33	0.00	0.63	0.46	**0.51**
Esp.4	*Precision*	0.71	0.53	0.63	0.00	0.57	**0.61**
	Recall	0.55	0.11	0.21	0.00	0.23	**0.27**
Esp.5	*Precision*	0.23	0.60	0.46	0.23	0.00	**0.38**
	Recall	0.43	0.30	0.37	0.57	0.00	**0.42**

Observamos que tanto o *precision* quanto o *recall* alcançados pela abordagem (Tabela 2 e Tabela 3) são próximos aos valores obtidos pelos mapas dos especialistas (Tabela 4 e Tabela 5). Por meio deste experimento, podemos observar que a atividade de construção de mapas a partir de textos é complexa e subjetiva, mesmo para especialistas, o que ressalta a dificuldade de construir automaticamente um mapa que represente o conhecimento consensual sobre o domínio de um texto.

Por fim, verificamos que embora o valor alcançado por nossa abordagem ainda não seja suficiente, apenas 16 conceitos (destacados em azul na Figura 8) dos 53 conceitos que compõem o mapa construído automaticamente não foram representados em um dos mapas construídos pelos especialistas.

9 Trabalhos Relacionados

Uma revisão da literatura sobre abordagens tecnológicas para construção automática de mapas conceituais a partir de textos, entre os anos de 1994 e 2016, nas bases *IEEE Xplore, ACM* e *Elsevier Science Direct* originou uma categorização a fim de obter uma análise objetiva sobre essas abordagens, detalhado em (Aguiar, Cury, & Zouaq, 2018). Assim, aplicamos a categorização às pesquisas recuperadas na revisão da literatura selecionando objetivamente as pesquisas com características a fins à apresentada neste artigo.

A abordagem de Wang, Cheung, Lee, & Kwok (2008) gera mapas conceituais a partir de resumos em inglês. Utiliza análise morfológica e sintática, identificando os elementos com base na estrutura das frases e regras sintáticas. Aplica a normalização para corrigir erros ortográficos, depende da detecção de sinônimos e da resolução anáfora. Usa análise estatística para verificar a relevância das proposições, fazendo uso da interação com o usuário para definir proposições incertas.

A abordagem de Zubrinic, Kalpic, & Milicevic (2012) gera mapas a partir de documentos legais em língua croata como um resumo do texto. Cria mapas hierárquicos de uma área específica usando o thesaurus do domínio. A partir de um corpus de domínio, os documentos são pré-processados e os metadados são mapeados. Usa técnicas linguísticas para lematização, reconhecimento de entidades, resolução de co-referência, análise léxica e sintática. Os conceitos são identificados a partir dos metadados e da frequência dos termos no texto. As proposições são extraídas a partir do padrão sujeito-predicado-objeto que contenha os conceitos identificados e relações estabelecidas entre os conceitos em um thesaurus. Uma estrutura de árvore formada por 25-30 conceitos das proposições é construída hierarquicamente atribuindo o título do texto como nó raiz.

A abordagem de Zouaq & Nkambou (2009) gera mapas conceituais de textos em inglês como etapa intermediária para gerar uma ontologia. Para isso, utiliza técnicas linguísticas de segmentação, normalização, análise estatística e sintática. Aplica o aprendizado de máquina para identificar palavras-chave e cria um mapa conceito semântico de frases contendo essas palavras-chave. As triplas são extraídas a partir das regras sintáticas e dependências gramaticais entre as palavras na frase. Os padrões léxico-semântico interpretam essa estrutura para extrair conceitos e relações. Finalmente, realiza análise estatística para definir a relevância de conceitos e relações.

A abordagem de De La Villa, Aparicio, Maña, & De Buenaga (2012) gera mapas conceituais de texto clínico em língua inglesa. Esta abordagem usa conceitos e uma ontologia para obter ricas informações sobre o domínio. O sistema pré-processa um conjunto de termos médicos compilados em uma lista e busca por termos do domínio no texto. O usuário escolhe

um conceito e consultas são realizadas na base de conhecimento para recuperar informações sobre o conceito.

A Figura 9 apresenta os mapas gerados pelas abordagens discutidas acima. Analisando as abordagens e seus respectivos mapas, podemos observar: a Figura 9 (a) (Wang, Cheung, Lee, & Kwok, 2008) apresenta mapa fragmentado em porções e rótulo de conceitos longo ou formado por pronome ou preposição; a Figura 9 (b) (Zubrinic, Kalpic, & Milicevic, 2012) apresenta mapa com rótulo de relação ausente, utiliza o thesaurus como fontes de dados além do texto e o mapa representa apenas um domínio específico; a Figura 9 (c) (Zouaq & Nkambou, 2009) apresenta mapa criado a partir de um conjunto de documentos e utiliza ontologia como fonte de dados além do texto; e a Figura 9 (d) (De La Villa, Aparicio, Maña, & De Buenaga, 2012) apresenta mapa criado a partir de um pequeno texto contendo algumas sentenças, utiliza base de conhecimento como fonte de dados além do texto e representa apenas um domínio específico.

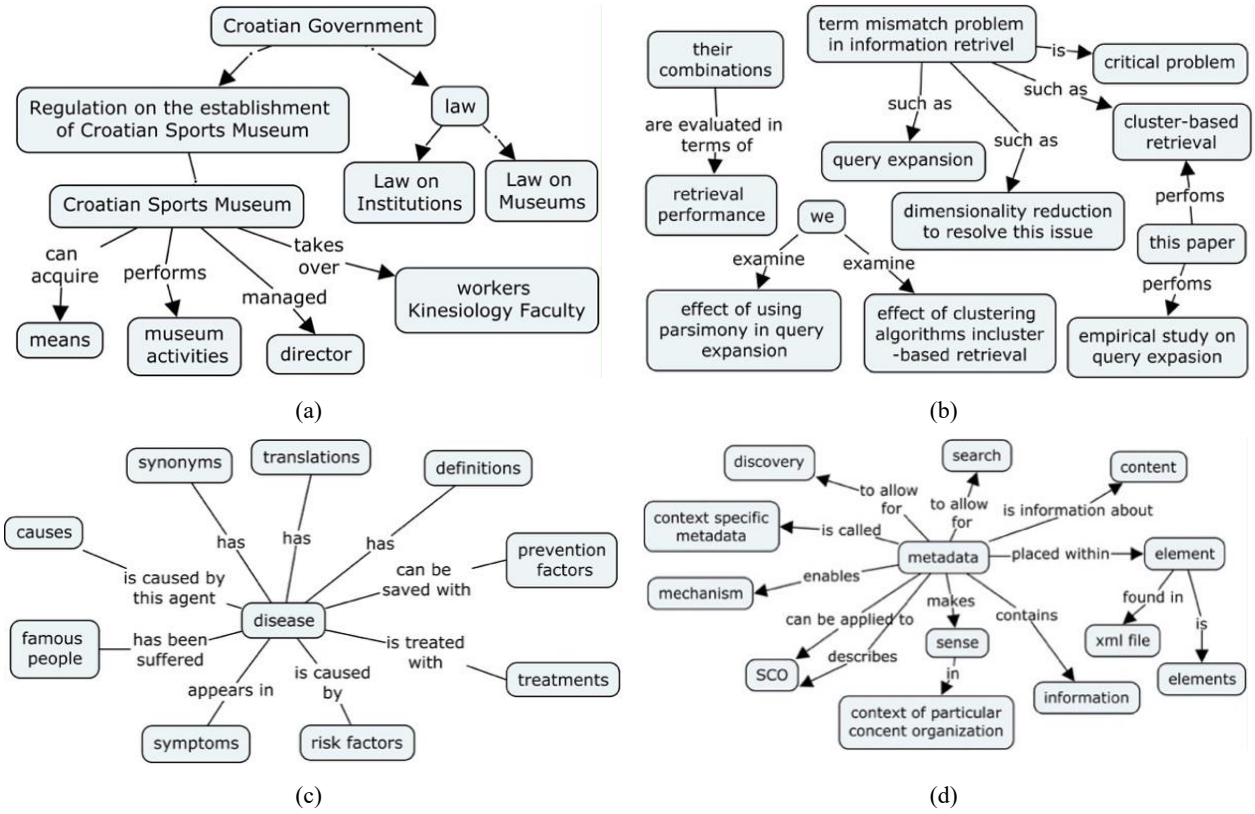

Figura 9: Mapas conceituais construídos pelos trabalhos relacionados

Realizando uma análise subjetiva entre o mapa construído por nossa abordagem (Figura 8) e os mapas apresentados pelos trabalhos relacionados (Figura 9), notamos alguns pontos fortes associados ao mapa construído por nossa abordagem:

- Todos os conceitos são conectados por frases de ligação, não havendo fragmentos ou conceitos livres;
- Rótulos são diretamente extraídos a partir da fonte de dados;
- Rótulo de conceitos são pequenos, não constituídos por pronomes e formados por muti-words, quando aplicável;
- Rótulo de relações são significantes e formados por verbos, algumas vezes não mencionado explicitamente no texto;
- Conceitos e proposições não exibem redundância.

No entanto, notamos alguns pontos fracos, tais como:

- Entidades nomeadas são usadas como rótulos para conceitos;
- Informação importante do texto tem sido perdida na sumarização;
- Lematização de conceitos prejudicou a compreensão das proposições.

Ademais, destacamos como principais contribuições da abordagem proposta, o uso do idioma Português, uma vez que os trabalhos relacionados usam idioma Inglês e Croata; a proposta de um método de mineração de mapas conceituais contemplando todas as etapas de processamento do texto desde a fonte de dados até a visualização e o desenvolvimento de uma abordagem direcionada a sumarização de textos em mapas conceituais.

10 Considerações Finais

Dado o desafio presente na construção automática de mapas conceituais a partir de textos, apresentamos neste artigo uma nova abordagem baseada em técnicas linguísticas aplicada a textos em idioma português. A abordagem apresentada se difere dos trabalhos relacionados tanto pela cobertura da proposta, desde a definição da fonte de dados até a visualização do mapa, quanto pelos resultados alcançados, comparando o mapa sumarizado da nossa abordagem com os pequenos mapas das abordagens relacionadas. Como já discutido, a abordagem produz resultados satisfatórios, mas ainda requer algumas melhorias, principalmente no que diz respeito a resolução de anáfora, lematização, analise sintática e análise semântica.

Nossa abordagem introduz a ideia de sumarização como papel fundamental para a representação do texto em mapa conceitual, visando superar as limitações observadas nas abordagens relacionadas. Um texto, mesmo com poucas sentenças, torna-se um mapa conceitual ilegível e incompreensível se for composto por todas as proposições existentes no texto. Assim, uma abordagem que considere a relevância dos conceitos para o domínio do texto tende a criar mapas conceituais mais significativos e que representem a sumarização do texto.

O artigo introduz o CMBuilder, ferramenta para construção automática de mapas conceituais sumarizados a partir de textos, desenvolvida para ambiente web e acesso público cujo objetivo principal é auxiliar o aprendizado de textos científicos. A ferramenta esta sendo incorporada a um portal de aprendizagem e incluída em uma arquitetura pedagógica baseada em mapas conceituais que será aplicada em nossa universidade. Ademais, o artigo também introduz o ExtroutMap, api que provê recursos computacionais para processamento, manipulação e extração de mapas conceituais. Para demais interessados, ExtroutMap pode ser utilizado como recurso para o desenvolvimento de novas ferramentas relacionadas a mapas conceituais.

Trabalhos futuros estão direcionados à melhoria da abordagem e dos recursos de processamento de linguagem natural para o português, bem como a expansão da api ExtroutMap com a integração de novos serviços.

Referências

Afonso, S., Bick, E., Haber, R., & Santos, D. (2002). Floresta sintá(c)tica: a treebank for Portuguese. *LREC'2002*, (pp. 1698-1703). Paris. Retrieved from https://www.linguateca.pt/documentos/AfonsoetalLREC2002.pdf [GS SEARCH]

Aguiar, C. Z., & Cury, D. (2017). Mineração de Mapas Conceituais a partir de Textos em Português. *Simpósio Brasileiro de Informática na Educação-SBIE*, *28.* Retrieved from *http://www.br-ie.org/pub/index.php/sbie/article/view/7640* doi: 10.5753/cbie.sbie.2017.1117. [GS SEARCH]

Aguiar, C. Z., Cury, D., & Zouaq, A. (2016). Automatic Construction of Concept Maps from Texts. *Conference on Concept Mapping - CMC.* Tallinn. Retrieved from http://cmc.ihmc.us/cmc2016papers/cmc2016-p90.pdf [GS SEARCH]

Aguiar, C. Z., Cury, D., & Zouaq, A. (2017). Mineração de Mapas Conceituais para Sumarização de Textos. *Workshops do Congresso Brasileiro de Informática na Educação, 6.* Retrieved from http://www.br-ie.org/pub/index.php/wcbie/article/view/7364 doi: 10.5753/cbie.wcbie.2017.57 [GS SEARCH]

Aguiar, C. Z., Cury, D., & Zouaq, A. (2018). Towards Technological Approaches for Concept Maps Mining from Text. *CLEI Electronic Journal, 21*, 7. Retrieved from http://www.clei.org/cleiej-beta/index.php/cleiej/article/view/261 doi: 10.19153/cleiej.21.1.7

Branco, A. e. (2010). Developing a Deep Linguistic Databank Supporting a Collection of Treebanks: the CINTIL DeepGramBank. In: LREC. Retrieved from https://pdfs.semanticscholar.org/b5d4/f4b2ddc6d2110e975e91c4233e458f1d8fde.pdf [GS SEARCH]

Chen, N.-S., Wei, C.-W., & Chen, H.-J. (2008). Mining e-Learning domain concept map from academic articles. *50*, 1009-1021. Retrieved from https://www.sciencedirect.com/science/article/pii/S0360131506001497 doi: 10.1016/j.compedu.2006.10.001 [GS SEARCH]

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. *ACM international conference on Intelligent User Interfaces.* ACM. Retrieved from https://dl.acm.org/citation.cfm?id=2166978 doi: 10.1145/2166966.2166978 [GS SEARCH]

Fonseca, E. R., & Rosa, J. L. (2013). Mac-Morpho revisited: Towards robust part-of-speech tagging., (pp. 98-107). Retrieved from http://www.aclweb.org/anthology/W13-4811 [GS SEARCH]

Gava, T. B., Menezes, C. d., & Cury, D. (2003). Aplicações de mapas conceituais na educação como ferramenta metacognitiva. *II International Conference on Engineering and Computer Education-ICECE.* Retrieved from http://www.geografia.fflch.usp.br/posgraduacao/apoio/apoio_raffo/flg5052/aula_1/AplicacoesdeMapasconceituaisnaEducacao.pdf [GS SEARCH]

Hutchins, J. (1987). Summarization: Some problems and methods. *Meaning: The frontier of informatics.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.104.9979 doi: 10.1.1.104.9979 [GS SEARCH]

Karannagoda, E. L., Herath, H. M., Fernando, K. N., Karunarathne, M. W., De Silva, N. H., & Perera, A. S. (2013). Document analysis based automatic concept map generation for enterprises. *Advances in ICT for Emerging Regions (ICTer).* IEEE. Fonte: Advances in ICT for Emerging Regions (ICTer). Retrieved from https://ieeexplore.ieee.org/document/6761171 doi: 10.1109/ICTer.2013.6761171 [GS SEARCH]

Le Coadic, Y.-F. (1996). *A ciência da informação.* Briquet de lemos Livros.

Leake, D., Maguitman, A., & Reichherzer, T. (2004). Understanding knowledge models: Modeling assessment of concept importance in concept maps. *Proceedings of the 26th*

conference CSS. Retrieved from https://pdfs.semanticscholar.org/36a9/e6fd7a154dfe0d74b5e2cd2e15681491e1e4.pdf [GS SEARCH]

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & Mcclosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. (pp. 55-60). ACL (System Demonstrations). Retrieved from https://nlp.stanford.edu/pubs/StanfordCoreNlp2014.pdf [GS SEARCH]

Mcgarry, K., & De Lemos, H. V. (1999). *O contexto dinânico da informação: uma análise introdutória.* Briquet de Lemos.

Novak, J. D., & Cañas, A. J. (2008). The theory underlying concept maps and how to construct and use them. Retrieved from https://web.stanford.edu/dept/SUSE/projects/ireport/articles/concept_maps/The Theory Underlying Concept Maps.pdf [GS SEARCH]

Novak, J. D., & Cañas, A. J. (2010). A teoria subjacente aos mapas conceituais e como elaborá-los e usá-los. *5*, pp. 9-29. Práxis Educativa. Retrieved from http://www.revistas2.uepg.br/index.php/praxiseducativa/article/view/1298 doi: 10.5212/PraxEduc.v.5i1.009029 [GS SEARCH]

Thakkar, K. S., Dharaskar, R. V., & Chandak, M. B. (2010). Graph-based algorithms for text summarization. *Emerging Trends in Engineering and Technology (ICETET).* IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/5698380 doi: 10.1109/ICETET.2010.104 [GS SEARCH]

Toffler, A. (1970). Future shock. New York: Amereon Ltd.

Villalón, J. J., & Calvo, R. A. (2011). Concept Maps as Cognitive Visualizations of Writing Assignments. *Educational Technology & Society, 14*, 16-27. Retrieved from https://www.j-ets.net/ets/journals/14_3/3.pdf [GS SEARCH]

Wang, W. M., Cheung, C. F., Lee, W. B., & Kwok, S. K. (2008). Mining knowledge from natural language texts using fuzzy associated concept mapping. . *Information Processing & Management, 44*, 1707-1719. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0306457308000526 doi: 10.1016/j.ipm.2008.05.002 [GS SEARCH]

Zouaq, A., & Nkambou, R. (2009). Evaluating the generation of domain ontologies in the knowledge puzzle project. *IEEE Transactions on Knowledge and Data Engineering*. Retrieved from https://ieeexplore.ieee.org/abstract/document/4752828 doi: 10.1109/TKDE.2009.25 [GS SEARCH]

Zubrinic, K., Kalpic, D., & Milicevic, M. (2012). The automatic creation of concept maps from documents written using morphologically rich languages. *Expert systems with applications, 39*, 12709-12718. Retrieved fromhttps://dl.acm.org/citation.cfm?id=2343262 doi: 10.1016/j.eswa.2012.04.065 [GS SEARCH]