Agrupando estudantes a partir da similaridade semântica de mapas conceituais

Rodrigo Ruy Boguski[1], Davidson Cury[1]

[1]Departamento de Informática – Universidade Federal do Espírito Santo (UFES)
Vitória – ES – Brasil

{rodrigoboguski, dedecury}@gmail.com

Abstract. *Semantically comparing content produced by different students and identifying existing clusters based on this similarity is a major challenge for teachers. This work presents a proposal for the formation of semantic clustering of students from the semantic comparison of conceptual maps constructed by them. Its approach is the automated reading of concept maps, using them as input to vector models of natural language processing that consider thematic and semantic aspects and the context of words. The analyzes carried out make it possible to compare the conceptual models that different individuals have on a subject and plan interactions between them.*

Resumo*. Comparar semanticamente conteúdos produzidos por diferentes alunos e identificar grupos existentes com base nessa semelhança é um grande desafio para os professores. Este trabalho apresenta uma proposta para a formação de agrupamentos semânticos de alunos a partir da comparação semântica de mapas conceituais por eles construídos. Sua abordagem é a leitura automatizada de mapas conceituais, utilizando-os como entrada para modelos vetoriais de processamento de linguagem natural que consideram aspectos temáticos, semânticos e o contexto das palavras. As análises realizadas permitem comparar os modelos conceituais que diferentes indivíduos possuem sobre um assunto e planejar interações entre eles.*

1. Introdução

O estabelecimento de uma comunicação efetiva entre os participantes de um processo de aprendizagem é um desafio, pois, para que isso aconteça, é necessário que além do conhecimento dos símbolos utilizados na linguagem corrente, também se conheça sua sintaxe e semântica. A sintaxe se deve às regras estruturais da linguagem e, uma vez definidas, obedecem a um certo rigor morfológico que facilita a comunicação, pois, para estabelecê-la de maneira literal, é preciso apenas que os participantes obedeçam ao contrato linguístico. A semântica depreende esforço maior porque está vinculada a um conjunto de parâmetros, tais como contexto, pontuação e encadeamento entre as palavras, que, quando alterados, mudam objetivamente o significado do conteúdo transmitido. Quando olhamos para o campo educacional, surge uma série de aplicações com necessidades de avaliação semântica, tais como, a identificação de questões semelhantes em fóruns disciplinares, a atribuição de notas próximas para respostas semelhantes a uma mesma questão, ou a formação de grupos que representam consistência semântica em termos de compreensão de um assunto por seus membros. Todas essas situações têm como questão central a verificação da semelhança entre conteúdos e buscam garantir que situações semanticamente semelhantes sejam tratadas da mesma forma em todos os casos. Assim, a correta verificação de semelhanças visa promover uma medida de tratamento.

DOI: 10.5753/sbie.2022.224957

Quando os conteúdos são representados na forma textual, uma possível verificação de similaridade que pode ser determinada é quão próximos eles estão sintaticamente. No entanto, ainda que uma análise sintática, em que sejam consideradas exclusivamente as palavras presentes nos textos, seja relevante para determinar o tema do discurso (Kim & Gil, 2019), não é suficiente para determinar seu contexto semântico. Dessa forma, é necessário verificar como se dá a relação posicional de uma palavra com as palavras que a cercam, visto que, geralmente, a diferença em sua ordem está intrinsecamente ligada à mudança de sentido do texto.

Temos trabalhado há décadas com o objetivo de permitir que os aprendizes sejam capazes de representar sua estrutura cognitiva de forma simples e concreta, usando mapas conceituais em diferentes contextos (Boguski, Cury, & Gava, 2019), (Aguiar, Cury, & Zouaq, 2017), (Rios, Aguiar, & Cury, 2017), (Boguski & Cury, 2018), principalmente em sala de aula (Boguski & Cury, 2019), (Moreira, Boguski, & Cury, 2021) e têm funcionado muito bem para comunicação, representação e construção do conhecimento, sem, no entanto, deixar de considerar os fatores citados anteriormente para uma comunicação eficaz e representação do conhecimento. A nossa preocupação, neste momento, é a identificação dos grupos semânticos aos quais os alunos podem pertencer e, a partir daí, promover melhores interações entre alunos de mesmos grupos ou grupos adjacentes com base nas características da sua proximidade. Esses grupos são constituídos pela semelhança que os conteúdos semânticos representados em seus mapas conceituais possuem e, consequentemente, pela compreensão que os alunos têm sobre o conhecimento do assunto neles descrito. Neste trabalho, é apresentado um framework desenvolvido para o reconhecimento e comparação semântica de conteúdos produzidos por diferentes alunos na forma de mapas conceituais. Essa comparação visa criar agrupamentos de alunos, construídos a partir das semelhanças encontradas em dois níveis: temático e semântico. A análise dos agrupamentos e seus integrantes permitirá detectar dispersões na comunicação, perceber as semelhanças e diferenças cognitivas entre diferentes grupos de alunos, assim como entre alunos, individualmente, além de auxiliar o professor no planejamento de interações significativas entre os alunos.

Este trabalho está organizado da seguinte forma: na Seção 2 apresentamos o framework proposto; na Seção 3 os resultados obtidos a partir de experimentos e na Seção 5, a conclusão e implicações para a prática docente.

2. Framework

O framework proposto utiliza técnicas de processamento de linguagem natural e algoritmos de redes neurais que modelam mapas conceituais como documentos de texto, representando-os como vetores multidimensionais para realizar operações de comparação e agrupamentos temáticos e semânticos. Ele é dividido em 4 etapas: Leitura de mapas conceituais, Normalização de texto, Agrupamento temático e Agrupamento semântico.

2.1. Leitura de mapas conceituais

Os mapas conceituais são uma ferramenta gráfica para representar o conhecimento na forma de grafos. Eles foram propostos por Novak (Novak & Gowin, 1984) e estruturam graficamente conceitos e relações na forma de proposições. Escritos em linguagem natural e de maneira hierárquica, permitem explicitar de forma objetiva quais conceitos estão relacionados a outros conceitos e aqueles que são mais importantes na percepção do aprendiz, sendo, inerentemente, filtrados conceitos menos relevantes, e desenhados

aqueles que possuem relacionamentos com valor semântico quando associados a outros conceitos, pelo menos do ponto de vista do aluno e de seu conhecimento do assunto até então. Assim, um mapa conceitual é um conjunto sintético de conceitos semanticamente relevantes, que podem, por exemplo, ter se originado de um texto em um artigo, ou mesmo livremente, a partir do conhecimento do aprendiz. Embora sua representação diagramática permita uma leitura e compreensão mais rápidas quando comparadas a um texto que possa tê-lo originado, muitas vezes ocorrem erros em sua construção que podem dificultar a compreensão das informações descritas. Por exemplo, caminhos em *loop*, ambiguidades conceituais, estruturação não hierárquica ou ainda, uso confuso de frases inteiras em vez de proposições. Embora isso seja suficiente para atender o critério de representação do conhecimento, gera muitas dificuldades de compreensão, quer seja para uma pessoa ou mesmo para processamento computacional. A representação conceitual, por meio de mapas, pelos alunos, por mais simples que seja, pode ser um pouco confusa até que um certo nível de qualidade seja alcançado. Assim, visando formar grupos de alunos que apresentem semelhança semântica, buscou-se uma estratégia que pudesse superar essas dificuldades, sem, no entanto, exigir um treinamento extensivo, dos alunos, para construir os mapas, permitindo que alunos construam mapas melhores a seu tempo.

Existem alguns trabalhos relacionados para leitura e comparação de mapas conceituais (Lamas, Boeres, Cury, & Menezes, 2005), (Marcos P. D. Lovati, 2017), (Limongelli, Sciarrone, Lombardi, Marani, & Temperini, 2017), contudo, residem na comparação das proposições extraídas da leitura do mapa de forma individual ou pela formação de n-gramas (Caldas & Favero, 2009). A proposta aqui apresentada compara mapas de forma mais ampla, considerando o conjunto de conceitos e relações, sua frequência de ocorrência e o relacionamento com os outros conceitos. Isso gera uma correspondência com espectro mais abrangente que a comparação de pares de proposições, resultando na explicitação natural de um assunto. No processo de leitura, o mapa é percorrido e extraído seu conteúdo, representando-o como sentenças textuais. O método para geração do texto a partir do mapa é iniciado com a identificação de nós iniciais e finais. Os nós iniciais são aqueles que não possuem arestas de saída. Nós finais são aqueles que possuem apenas arestas de entrada. Nós intermediários são aqueles que possuem arestas de entrada e saída. A Figura 1 explicita estes três tipos de nós em azul, verde e vermelho, respectivamente como, nó inicial, intermediário e final. Os nós iniciais são, geralmente, conceitos de alta hierarquia funcionando como subsunçores, ou âncoras, para outros conceitos (Ausubel, Novak, & Hanesian, 1978). Um caminho percorrível entre um nó inicial e nó final é obtido a partir do produto cartesiano P (nó inicial, nó final) que forneça um caminho contínuo e possua direção do nó inicial para o final, sem passar por *loops* no grafo. Esse método permite descrever a informação no mapa como um texto, semelhante ao que, provavelmente, o tenha originado, constituindo, no entanto, uma versão essencial, contendo apenas conceitos e associações relevantes.

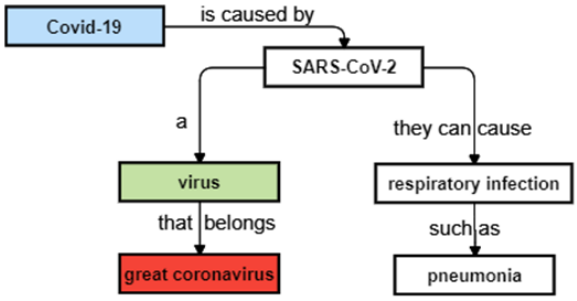

Figura 1. Tipos de nós conceituais

2.2. Normalização de texto

Um documento de texto é um objeto do tipo sequência de texto. Um *corpus* é definido como uma coleção de documentos. A normalização de texto ocorre usando técnicas de processamento de linguagem natural a fim de torná-lo adequado ao processamento por modelos computacionais. Nele, cada documento é submetido a um conjunto de transformações para torná-los adequados ao processamento de similaridade. São elas:
Filtragem: Pré-processar o documento para remover caracteres especiais como aspas duplas, quebras de linha e também padrões que não agregam valor semântico.
Sumarização: Extração das frases mais importantes de documentos. É baseado no algoritmo TextRank (Mihalcea & Tarau, 2004) e Barrios (Barrios, López, Argerich, & Wachenchauzer, 2015) e é usado quando o documento de entrada é muito grande.
Tokenização: Divide documentos em palavras usando espaço como delimitador, realizando, neste momento, reconhecimento de entidade e associação sintática.
Formação de bigramas e trigramas: Verificar palavras que fazem sentido quando aparecem em pares ou trios. Por exemplo, a palavra "francês" tem um significado diferente do que quando aparece em "Revolução Francesa".
Remoção de *stop words*: *stop words* são palavras que aparecem com frequência em um *corpus*, mas não agregam valor em termos de significado, como artigos e proposições, podendo causar distorções na hora de treinar o modelo, devendo ser removidas.
Lematização: Processo de conversão de uma palavra em sua forma base, semelhante ao processo de *stemming*. A diferença entre *stemming* e lematização é que esta considera o contexto e converte a palavra em sua forma base significativa, enquanto aquele remove apenas os últimos caracteres, o que pode levar a significados incorretos e erros de ortografia. Por esse motivo, optamos pela lematização em vez de *stemming*.

Verificamos que a ausência de uma dessas etapas ou a inversão de sua ordem altera negativamente a qualidade dos resultados em relação aos obtidos usando a sequência proposta. Além disso, as três últimas etapas foram as que apresentaram impacto mais substancial nos resultados, assim como adotar lematização em vez de *stemming*.

2.3. Agrupamento temático

Para realizar o agrupamento semântico é preciso de uma métrica de comparação semântica que permita verificar a semelhança de dois conteúdos quanto ao seu significado. Por exemplo, quando comparamos as frases "o gato comeu o rato" com "o rato comeu a comida do gato", poderíamos dizer, inicialmente, que elas estão próximas se estivéssemos olhando apenas a composição em termos das palavras presentes, porém, eles não são semanticamente iguais. Um passo antes da comparação semântica e com granularidade um pouco maior, é poder diferenciar se duas informações pertencem ao mesmo assunto ou tópico, pois, se não pertencerem, por mais vigorosos que sejam os esforços comparativos, não haverá acordo, tampouco aproximação semântica, dentro de diferentes temas. Assim, a verificação de similaridade temática é uma medida importante, pois consegue apontar, antes de uma verificação semântica, possíveis concordâncias e discordâncias que alunos possam ter sobre um tema. Além disso, possui uma complexidade inferior em relação à comparação semântica e isso a torna viável como análise preliminar. Embora não permita verificar se os participantes de uma conversa estão dizendo a mesma coisa, pode indicar, por outro lado, que falam da mesma coisa, e isso é relevante, por exemplo, para a atuação de um mediador na resolução de divergências entre eles. Por exemplo, considere as seguintes frases textuais extraídas de notícias sobre o assunto (tópico) COVID-19 (para Doença de Coronavírus).

Sentence 1: *Covid-19 disease, which appeared in Wuhan, China in late 2019, is caused by SARS-CoV-2, a virus that belongs to the great coronavirus. Very frequently, they can cause a simple cold as well as a serious respiratory infection such as pneumonia, causing fatal epidemics as was the case with SARS or MERS and now with Covid-19 (for Coronavirus Disease).*

Sentence 2: *The world has seen the emergence of a Novel Corona Virus, caused by SARS-CoV-2, on 31 December 2019, officially referred to as COVID-19. The virus was first isolated from persons with pneumonia in Wuhan city, China. The virus can cause a range of symptoms, ranging from mild illnesses like cold to pneumonia. Symptoms of the disease are fever, cough, sore throat, and headaches.*

Sentence 3: *Hydroxychloroquine is currently being studied for the treatment and prevention of coronavirus disease 2019 (COVID-19), caused by SARS-CoV-2. Only limited clinical study information is available at this time to support the use of hydroxychloroquine for the treatment or prevention of COVID-19.*

Na análise das três sentenças acima quanto à similaridade, verificou-se que todas elas possuem uma aproximação temática, podendo ser agrupadas em um mesmo agrupamento temático (COVID-19). No entanto, apenas os dois primeiros podem formar um agrupamento semântico que representa a mesma informação descrita. O terceiro, por não possuir o mesmo valor semântico dos demais, deve ser categorizado em um novo grupo semântico. Quando se pensa em agrupamento semântico, pretende-se que as correspondências entre sentenças sejam sensíveis a essa variação contextual. Para satisfazer esses requisitos, foi utilizada uma abordagem que representasse os documentos de texto extraídos dos mapas, como vetores N-dimensionais de características que exploram aspectos quanto à aderência temática e semântica, considerando para isto, a composição gramatical do texto e o encadeamento entre as palavras. Com base nessa representação, foram feitas comparações usando modelos que utilizam uma medida da distância entre essas características. Para realizar a verificação temática, utilizou-se como suporte o modelo LDA ou *Latent Dirichlet Allocation* (Blei, Ng, & Jordan, 2003). É um modelo probabilístico generativo para coleções de documentos e apresenta uma forma de descobrir automaticamente a probabilidade de um documento pertencer a determinados temas (também chamados de tópicos) a partir da análise probabilística da ocorrência de palavras nele contidas. A ideia básica é que os documentos sejam representados como misturas aleatórias em tópicos latentes, em que cada tópico é caracterizado por uma distribuição de palavras. Do ponto de vista associativo, palavras relacionadas a um determinado tema tendem a aparecer com certa associação (Boguski & Cury, 2018).

O modelo LDA é um algoritmo que utiliza uma rede neural, recebendo como entrada um *corpus* e a quantidade k de tópicos que se espera que ele contenha. Em seguida, procede-se à reorganização temática com a distribuição de cada documento e respectivas palavras-chave para cada tópico, procurando obter uma boa composição da distribuição das palavras-chave por tópico. Um tópico nada mais é do que uma coleção de palavras-chave dominantes que normalmente o representam. Assim, apenas olhando as palavras-chave do tópico (tema), podemos identificar do que se trata. Uma boa segregação de tópicos considera fatores como a qualidade do processamento de texto (normalização), a variedade de tópicos sobre os quais o texto fala, a escolha ajustada dos parâmetros da rede neural e a quantidade de tópicos latentes. Para obter, automaticamente, a quantidade ideal de temas (tópicos) esperado na entrada do modelo, foi utilizada a técnica de análise da evolução do valor de coerência de tópico (Newman, Lau, Grieser,

& Baldwin, 2010), obtendo-o iterativamente, para cada grupo temático k, em que k=2...N (conjunto máximo). O melhor valor de k encontrado é aquele que possui o maior valor de coerência de tópico. Uma limitação ao usar o modelo LDA é que nem sempre obtemos os mesmos resultados comparativos para os mesmos valores de entrada, pois é um modelo probabilístico. Para reduzir essas variações, temos trabalhado para obter melhores resultados de ajustes nos parâmetros da rede neural, bem como melhorar a qualidade da formação de textos a partir de mapas conceituais.

2.4. Agrupamento semântico

O agrupamento temático realizado com o LDA considera as palavras do texto, mas não analisa o relacionamento entre elas. Para solucionar essa deficiência e alcançar uma dimensão semântica, foi utilizado o modelo Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013) que cria vetores de representações de palavras, utilizando uma rede neural rasa, para capturar a semântica do documento. Ele realiza a correspondência de palavras usando um espaço vetorial menor e resulta em um conjunto de vetores de palavras. Vetores próximos no espaço vetorial têm significados semelhantes, com base no contexto, e vetores de palavras distantes têm significados diferentes. Por exemplo, os vetores gerados pelas palavras “forte” e “poderoso” estariam semanticamente próximos enquanto “forte” e “prédio” estariam relativamente distantes. A suposição subjacente do Word2Vec é que duas palavras que compartilham contextos semelhantes também compartilham um significado semelhante e, consequentemente, uma representação vetorial semelhante do modelo. Dessa forma, a verificação de similaridade é feita comparando os vetores obtidos para cada palavra em um documento. Para calcular um vetor para o documento como um todo, foi utilizado o algoritmo Doc2Vec (Le & Mikolov, 2014), uma adaptação do Word2Vec que geralmente supera uma média simples dos vetores gerados por este modelo. No Doc2Vec, os vetores de um documento são obtidos a partir da inferência, dada pelo treinamento de uma rede neural na tarefa sintética de prever uma palavra central com base na média dos vetores de palavras de contexto e do vetor de documentos completo do *corpus*, dando o resultado uma dimensão semântica.

Como o modelo Doc2Vec utiliza uma rede neural, é necessário treiná-la e testá-la quanto à assertividade para inferência de vetores. Durante a construção do nosso framework, foram testadas duas abordagens de treinamento, a fim de ver qual seria a mais adequada para o propósito. Na primeira abordagem, o modelo vetorial é treinado a partir de um corpus mais amplo, denominado text8, obtido pela limpeza de quase 5 GB de artigos da Wikipédia, contando cerca de 17 milhões de palavras sobre diferentes temas. A ideia principal era que o modelo pudesse ter conhecimento sobre diferentes assuntos e, assim, ter um escopo mais abrangente na tarefa de comparar possíveis temas que poderiam ser utilizados na construção de mapas conceituais. Quando o modelo treinado foi avaliado, verificou-se que, embora tivesse fornecido resultados positivos, apresentava algumas variações para inferência de novos vetores, tornando o processo de comparação de semelhanças inconsistente e posteriormente influenciando na formação de agrupamentos semânticos. Esse resultado ocorre quando o *corpus* de treinamento é muito diferente do *corpus* de teste, ou quando não possui amostras suficientes dele. Na segunda abordagem, foi utilizado para treinamento e teste, um *corpus* de treinamento privado, originado de mapas conceituais de diferentes atividades sobre o mesmo tema. Isso proporcionou excelentes resultados para inferência de novos vetores a partir do modelo, apresentando também uma consistência de validação cruzada do *corpus* frente ao

aprendizado do modelo para a rede neural. Essa segunda abordagem proporcionou melhores resultados comparativos porque os documentos utilizados pertenciam ao mesmo tema, reduzindo as variações ocorridas na primeira abordagem devido ao escopo mais amplo. Outra vantagem de seu uso é a construção e formação de um *corpus* temático específico, que é ampliado à medida que novos experimentos são realizados. Depois que o modelo foi treinado e os vetores para o *corpus* do mapa conceitual foram inferidos, realizou-se comparações de similaridade usando duas medidas: a similaridade do cosseno (Gan, Ma, & Wu, 2007) e a medida do cosseno suave (Sidorov, Gelbukh, Gomez-Adorno, & Pinto, 2014). A similaridade do cosseno (Singhal, 2001) é uma medida de similaridade entre dois vetores que compara o ângulo entre os vetores normalizados resultantes e a medida do cosseno suave é um método que permite avaliar a similaridade entre dois documentos de forma significativa, mesmo quando eles não têm palavras em comum. Ele usa uma medida de semelhança entre as palavras que pode ser derivada do casamento de palavras usando o Word2Vec, superando muitos dos métodos avançados na tarefa de semelhança semântica de textos (Charlet & Damnati, 2017), (Novotný, 2018).

Foram testados dois algoritmos de agrupamento para compor o framework na tarefa de realizar agrupamento semântico: K-Means e BIRCH (Gan, Ma, & Wu, 2007). Inicialmente, escolhe-se o K-Means (Kanungo, et al., 2002), no entanto, seu cálculo inicial do centroide é escolhido aleatoriamente, apresentando, muitas vezes, diferentes resultados para cada execução. Mesmo que essas diferenças sejam pequenas, é necessário executar o algoritmo várias vezes e escolher o resultado que produz os melhores grupos. Para pequenos grupos, ele é relativamente estável. O agrupamento BIRCH (Zhang, Ramakrishnan, & Livny, 1996) tem a vantagem de apresentar sempre o mesmo resultado, pois é gerado a partir da mesma função de distribuição em um processo de agrupamento hierárquico. Além disso, pode formar grupos de forma incremental e dinâmica para os dados recebidos. Na maioria dos casos, o BIRCH requer apenas uma única verificação do banco de dados. Por esta razão, o agrupamento BIRCH foi escolhido em vez de K-Means. Para avaliar a qualidade do agrupamento gerado, utilizou-se a análise do valor de silhueta (Saputra, Saputra, & Oswari, 2019) que mede a distância de separação entre os grupos resultantes, fornecendo uma medida de quão próximo cada ponto em um grupo está de pontos em grupos vizinhos. Complementarmente, foi usada a análise do gráfico do cotovelo (Saputra, Saputra, & Oswari, 2019) que considera a medida da soma dos quadrados das distâncias dos participantes do grupo ao centro do grupo mais próximo.

3. Resultados

O framework proposto foi aplicado em cerca de 450 testes laboratoriais e após verificar-se a eficácia, foram realizados experimentos supervisionados em sala de aula, em diferentes turmas do curso de ciência da computação e engenharia da computação em nossa universidade. Em cada atividade, cada aluno produziu um único mapa conceitual. O experimento escolhido para detalhamento de resultados foi realizado em uma turma de 9 alunos do curso de ciência da computação, resultando na produção de 9 mapas conceituais. Para cada aluno, foi distribuído um texto, extraído de notícias da internet sobre a pandemia do COVID-19. Assim, cada aluno ficou responsável por produzir um único mapa conceitual. O *corpus* resultante da leitura dos mapas é composto pelo seguinte conjunto de documentos: *Corpus* = {D1, D2, D3, D4, D5, D6, D7, D8, D9], onde, D1 representa o documento de texto extraído do mapa do aluno 1 e assim sucessivamente.

Os textos gerados foram submetidos a um grupo de 4 pessoas para que pudessem identificar, a partir de sua inspeção, quais os possíveis temas neles existentes. Desta análise, emergiu naturalmente a existência de 3 tópicos ou temas diferentes, são eles: 1) o início do surto do novo coronavírus e seus efeitos, 2) a eficácia da cloroquina no combate ao coronavírus, e 3) estudos sobre o uso da cloroquina em diferentes países. Observe que os temas identificados são muito próximos e pertencem a um contexto mais amplo, o COVID-19. Ainda assim, pretende-se avaliar se a verificação de similaridades temáticas é sensível a essas pequenas granularidades temáticas. Nessa perspectiva, foi atribuído o valor de k = 3 à quantidade de agrupamentos temáticos a serem considerados, para fins de validação de hipótese. Em seguida, procedeu-se ao cálculo da verificação de semelhanças e formação de agrupamentos temáticos. Os documentos D_x atribuídos a um tópico, assim como a sua probabilidade de aderência a ele, exibida entre parêntesis, foram tópico 1: D1(0,9997), D4(0,9998), D6(0,9998), D8(0,9997), tópico 2: D2(0,9998), D5(0,9997), D9(0,9998) e tópico 3: D3(0.9998), D7(0.9997). Para calcular a coerência do tópico, obtivemos os valores de coerência para cada k valor de tópico, variando k iterativamente do tamanho 1 (agrupamento mínimo) à quantidade de documentos N do *corpus* (Tabela 1).

Tabela 1. Valor de coerência de tópico por tópico

Nº Tópico	1	2	3	4	5	6	7	8	9
Coerência Tópico	0,341	0,378	0,458	0,411	0,402	0,356	0,219	0,199	0,150

O gráfico abaixo (figura 2), gerado a partir da Tabela 1, mostra que o valor de coerência aumenta com a quantidade de tópicos, havendo um declínio (1ª inflexão) entre a quantidade de 3 a 4 tópicos. Portanto, foi escolhido o pico do valor de coerência do tópico, ocorrendo para a quantidade de tópicos igual a 3, confirmando a escolha feita previamente por inspeção humana.

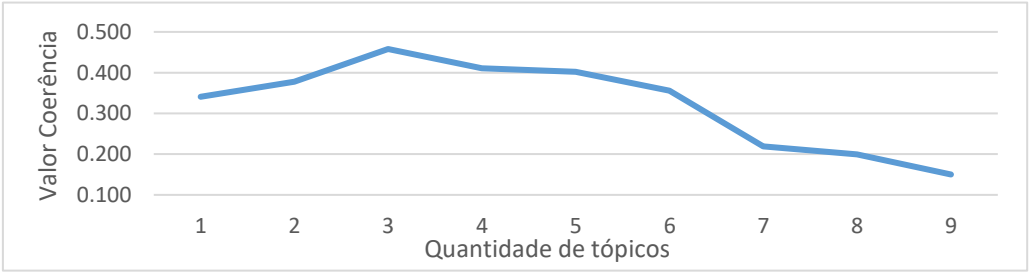

Figura 2. Distribuição da pontuação de coerência por tópico

Para realizar comparações semânticas, foram inferidos novos vetores do modelo Doc2Vec e comparados usando a similaridade de cosseno e a medida de cosseno suave. Constatou-se que as duas métricas apresentam resultados ligeiramente diferentes, sendo que o uso do cosseno suave (Tabela 2) produz melhores resultados de comparações semânticas, tendo valores mais próximos aos realizados pela validação humana nos mesmos documentos. Por exemplo, enquanto o cosseno de similaridade aponta que os documentos D1 e D2 possuem correspondência semântica de 0,7806, a medida do cosseno suave indica apenas 0,3768. Este último resultado corresponde às correspondências de similaridade que foram realizadas manualmente pela equipe para casos como esses. Outro resultado satisfatório é quando se realizou a validação cruzada, ou seja, quando se espera que a rede neural avalie, em termos de similaridade, um documento já visto anteriormente. Esses casos apresentam resultados próximos a 100%.

Tabela 2. Comparação entre documentos usando a medida de cosseno suave

	D1	D2	D3	D4	D5	D6	D7	D8	D9
D1	0,996	0,377	0,271	0,952	0,622	0,967	0,684	0,881	0,700
D2	0,377	0,998	0,955	0,442	0,858	0,177	0,926	0,294	0,993
D3	0,271	0,955	0,989	0,284	0,872	0,188	0,838	0,237	0,300
D4	0,952	0,442	0,284	0,999	0,396	0,978	0,619	0,552	0,824
D5	0,622	0,858	0,872	0,396	0,972	0,557	0,865	0,528	0,488
D6	0,967	0,177	0,188	0,978	0,557	1,000	0,596	0,555	0,822
D7	0,684	0,926	0,838	0,619	0,865	0,596	0,994	0,414	0,569
D8	0,881	0,294	0,237	0,552	0,528	0,555	0,414	1,000	0,396
D9	0,700	0,993	0,300	0,824	0,488	0,822	0,569	0,396	1,000

Isso significa que os dados de teste validam tanto o modelo quanto a estratégia de treinamento e aprendizado da rede neural usada em nosso framework. Por fim, foi realizada a formação do agrupamento semântico utilizando o algoritmo BIRCH, a partir da verificação das similaridades semânticas encontradas entre os vetores inferidos a partir dos documentos do mapa conceitual. Na Tabela 3, são apresentados os resultados dos agrupamentos semânticos e seus respectivos valores de silhueta.

Tabela 3. Agrupamento semântico com o algoritmo BIRCH

K	AGRUPAMENTO BIRCH	SILHUETA
2	[D1,D4,D6,D8], [D2,D3,D5,D9,D7]	0,52200
3	[D1,D4,D6,D8], [D9], [D2,D3, D5,D7]	0,46337
4	[D1,D4,D6,D8], [D5], [D9], [D2,D3,D7]	0,34033
5	[D1,D4,D6,D8], [D3], [D5], [D9], [D2,D7]	0,27116

Uma importante constatação foi a confirmação da similaridade temática a partir da similaridade semântica. Por exemplo, foi verificado que os documentos D1, D4, D6 e D8, pertencentes ao grupo temático 1, durante a etapa de verificação temática, pertencem também ao mesmo agrupamento semântico, confirmando sua similaridade em um nível mais profundo do que o temático. Dessa forma, esses documentos não só pertencem ao mesmo assunto, mas também representam a mesma informação, é claro, dentro das devidas proporções de proximidade semântica. Outra informação relevante é que, embora os documentos D2, D5 e D9 pertençam ao mesmo grupo temático (tópico 2), integram diferentes grupos semânticos para k> 2 (Tabela 3). São apresentados na Figura 3 o gráfico Elbow à esquerda, com a indicação da melhor quantidade de agrupamentos em corte transversal pontilhado e à direita, a representação gráfica dos grupos em torno de seus respectivos centroides.

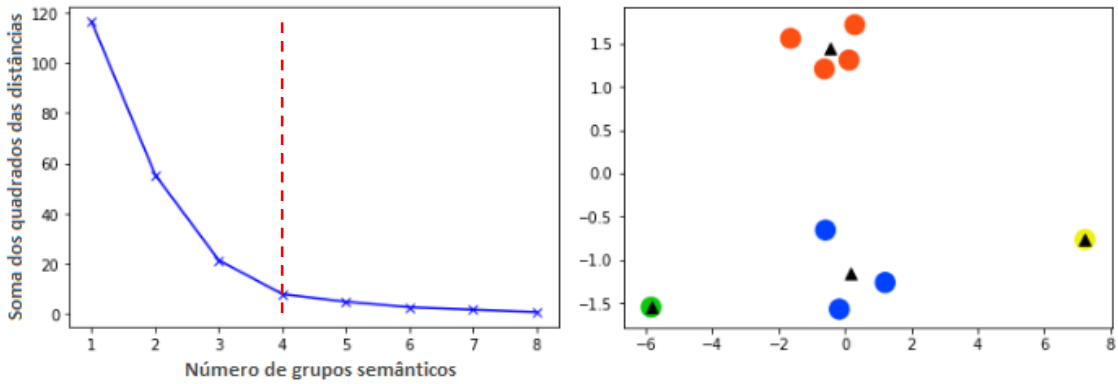

Figura 3. K ideal para **Elbow (esquerda) e a**grupamento semântico **(direita)**

Uma análise preliminar da silhueta (Tabela 3) em conjunto com uma análise complementar pelo método do cotovelo (Figura 3) permite refinar o entendimento. Existem, de fato, 4 agrupamentos, mas apenas 2 deles (grupos em azul e vermelho)

constituem grupos mais notáveis, já que os outros 2 grupos são formados por apenas 1 elemento cada. Isso confirma a indicação de 4 grupos pelo método do cotovelo, porém, o melhor agrupamento ocorre para k=2 (0,52200), conforme indicado pelo método da silhueta. A análise e compreensão do resultado dos diferentes indicadores são relevantes porque permitem evitar erros e também indicar, não só quais os alunos pertencem a um grupo, mas também medir a proximidade dos conteúdos por eles produzidos. Em relação ao grupo 1, por exemplo, formado pelos alunos {1,4,6,8}, é possível observar na Tabela 2, que o documento produzido pelo aluno 1 é mais semelhante semanticamente ao documento produzido pelo aluno 6 (96,66%) que o produzido pelo aluno 4 (95,15%) e pelo aluno 8 (88,06%). Isso permite planejar interações mais eficientes, levando em consideração quais grupos estão mais próximos e podem cooperar por meio de interações significativas ou ainda, de maneira mais refinada, quais alunos possuem maior similaridade semântica de conteúdo, e podem interagir, ainda dentro do mesmo grupo.

4. Conclusão e implicações para a prática docente

Neste trabalho, apresentamos um framework para a formação de agrupamentos semânticos de alunos a partir da leitura automatizada dos mapas conceituais por eles produzidos. Essa construção é feita, primeiramente, a partir de comparações temáticas e depois semânticas. A análise temática pode ser utilizada para buscar a divergência de contexto entre os alunos e esclarecer a necessidade da mediação do professor para resolver incongruências. A principal contribuição da análise semântica é a identificação da convergência semântica entre os alunos pela análise apenas de seus mapas conceituais e pode ser usada em uma variedade de aplicações educacionais, como a comparação de similaridade semântica de mapas conceituais que não são semelhantes em forma, atribuição de notas semelhantes para conteúdos semelhantes e formação de agrupamento semântico para atividades de sócio interações. A análise comparativa entre o texto original e o gerado pela leitura automatizada dos mapas conceituais, para formação de grupos semânticos, obteve uma correspondência acima de 80% entre os grupos realizados por inspeção humana, confirmando a viabilidade do método para fornecimento de suporte estruturante em interações entre grupos adjacentes ou mesmo, entre alunos de um mesmo grupo. Outra vantagem do framework é que ele permite que a comparação entre os mapas seja realizada diretamente a partir dos dados neles representados, dispensando assim o uso de técnicas de *scaffolding* que exigem o uso de metadados, originalmente utilizados em redes semânticas (Peters & Shrobe, 2003) ou representações de modelos ontológicos (Gómez-Gauchía, Díaz-Agudo, & Gonzalez-Calero, 2004), (Simón, Luigi, & Rosete, 2007). Contribui também como método de leitura automatizada de mapas conceituais sem que estes sejam rigidamente representados por proposições formais, aspecto bastante frequente, principalmente quando das primeiras experiências de construção de mapas.

Como continuação deste trabalho, pretendemos desenvolver um plano de interação por grupos proximais que considere as medidas de similaridades semânticas entre eles e entre cada aprendiz de um mesmo grupo, ambas apresentadas neste trabalho, para detectar também aproximações cognitivas na comunicação entre estudantes e auxiliar o professor no planejamento de interações significantes entre estudantes que pertençam a zonas de desenvolvimento proximal (ZDP) (Vygotsky L. S., 1978), (Vygotsky L. , 2007). Assim, integrantes de zonas de desenvolvimento proximais poderão cooperar mutuamente, realizando processos sociointeracionistas de ensino-aprendizagem que acreditamos ser capazes de promover aprendizagens de níveis superiores.

Referências

Aguiar, C. Z., Cury, D., & Zouaq, A. (2017). Mineração de Mapas Conceituais para Sumarização de Textos. *VI Congresso Brasileiro de Informática na Educação (CBIE 2017)*, (pp. 57-66).

Ausubel, D. P., Novak, J., & Hanesian, H. (1978). *Educational psychology: a cognitive view* (2nd ed.). New York: Holt Rinehart and Winston. doi:https://doi.org/10.1037/016814

Barrios, F., López, F., Argerich, L., & Wachenchauzer, R. (2015). Variations of the Similarity Function of TextRank for Automated Summarization. *Argentine Symposium on Artificial Intelligence.* Buenos Aires. Retrieved from https://arxiv.org/abs/1602.03606

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993-1022. Retrieved from https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Boguski, R. R., & Cury, D. (2018). Usando regras de associação para a identificação de falhas conceituais. *Simpósio Brasileiro de Informática na Educação (SBIE)*, (pp. 1443-1453). Fortaleza.

Boguski, R. R., & Cury, D. (2019). Fatores que influenciam a aprendizagem assistida em mapas conceituais. *XXX Simpósio Brasileiro de Informática na Educação.* Brasília.

Boguski, R. R., Cury, D., & Gava, T. (2019). TOM: An intelligent tutor for the construction of knowledge represented in concept maps. *Frontiers in Education (FIE).* Cincinnati.

Caldas, V. M., & Favero, E. L. (2009). Uma Ferramenta de Avaliação Automática para Mapas Conceituais como Auxílio ao Ensino em Ambientes de Educação a Distância. *XX Simpósio Brasileiro de Informática na Educação.* doi:http://dx.doi.org/10.5753/cbie.sbie.2009.%25p

Charlet, D., & Damnati, G. (2017). Soft-Cosine Semantic Similarity between Questions for Community Question Answering. *Proceedings of the 11th International Workshop on Semantic Evaluation*, (pp. 315–319). Vancouver, Canada. doi:http://dx.doi.org/10.18653/v1/S17-2051

Gan, G., Ma, C., & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications.* SIAM. doi:https://doi.org/10.1137/1.9780898718348

Gómez-Gauchía, H., Díaz-Agudo, B., & Gonzalez-Calero, P. A. (2004). Two-layered approach to knowledge representation using conceptual maps and description logics. *Concept Maps: Theory, Methodology, Technology, Proc. of the First Int. Conf. on Concept.* Retrieved from http://cmc.ihmc.us/Papers/cmc2004-205.pdf

Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., & Wu, A. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 881 - 892). doi:https://doi.org/10.1109/TPAMI.2002.1017616

Kim, S.-W., & Gil, J.-M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*. doi:https://doi.org/10.1186/s13673-019-0192-7

Lamas, F., Boeres, C., Cury, D., & Menezes, C. S. (2005). Comparando mapas conceituais utilizando correspondência de grafos. *Simpósio Brasileiro De Informática Na Educação - SBIE*, (pp. 24-27).

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on International Conference on Machine Learning*, (pp. 1188-1196). Retrieved from https://arxiv.org/abs/1405.4053

Limongelli, C., Sciarrone, F., Lombardi, M., Marani, A., & Temperini, M. (2017). A framework for comparing concept maps. *16th International Conference on Information Technology Based Higher Education and Training (ITHET).* doi:https://doi.org/10.1109/ITHET.2017.8067818

Marcos P. D. Lovati, C. Z. (2017). Clusterizando Mapas Conceituais para Identificar Desempenho Cognitivo em Grupos. *VI Congresso Brasileiro de Informática na Educação (CBIE 2017)*, (pp. 1397-1406). Recife, PE. doi:http://dx.doi.org/10.5753/cbie.sbie.2017.1397

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts. *Conference on Empirical Methods in Natural Language Processing*, (pp. 404–411). Barcelona, Spain. Retrieved from https://www.aclweb.org/anthology/W04-3252

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR Workshop Papers*. Retrieved from https://arxiv.org/abs/1301.3781

Moreira, R. B., Boguski, R. R., & Cury, D. (2021). Utilizando análise semântica para descobrir implicações significantes em mapas conceituais. *ANAIS DO XXXII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO*, (pp. 123-134). doi:https://doi.org/10.5753/sbie.2021.218489

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. *The 2010 Annual Conference of the In Human Language Technologies: North American Chapter of the Association for Computational Linguistics (NAACL-HLT '10)*, (pp. 100–108). Los Angeles, California. Retrieved from https://www.aclweb.org/anthology/N10-1012

Novak, J., & Gowin, D. (1984). *Learning how to learn.* Cambridge: Cambridge University Press. doi:https://doi.org/10.1017/CBO9781139173469

Novotný, V. (2018). Implementation Notes for the Soft Cosine Measure. *re. In Proceedings of the 27th ACM International Conference on Information and Knowledge Man*, (pp. 22-26). Torino, Italy. doi:http://doi.org/10.1145/3269206.3269317

Peters, S., & Shrobe, H. E. (2003). Using Semantic Networks for Knowledge Representation in an Intelligent Environment. *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003*, (pp. 323-329). doi:https://doi.org/10.1109/PERCOM.2003.1192756

Rios, P. T., Aguiar, C. Z., & Cury, D. (2017). Uma Abordagem construtivista para identificar o conhecimento usando mapa conceitual. *VI Congresso Brasileiro de Informática na Educação (CBIE 2017)*, (pp. 394-403).

Saputra, D. M., Saputra, D., & Oswari, L. D. (2019). Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method. *Sriwijaya International Conference on Information Technology and Its Applications.* doi:https://doi.org/10.2991/aisr.k.200424.051

Sidorov, G., Gelbukh, A., Gomez-Adorno, H., & Pinto, D. (2014). Soft Similarity and Soft Cosine Measure:Similarity of Features in Vector Space Model. *Computación y Sistemas*, 491–504. doi:https://doi.org/10.13053/cys-18-3-2043

Simón, A., L. C., & Rosete, A. (2007). Generation of OWL Ontologies from Concept Maps in Shallow Domains. *Congresos de la Asociación Española para la Inteligencia Artificial)(CAEPIA)*, (pp. 259-267). doi:https://doi.org/10.1007/978-3-540-75271-4_27

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bull IEEE Comput Soc Tech Comm Data Eng*, (pp. 35-43). Retrieved from http://www1.cs.columbia.edu/~gravano/cs6111/Readings/singhal.pdf

Vygotsky, L. (2007). A formação social da mente. In M. Fontes (Ed.), *Interação entre aprendizado e desenvolvimento* (7 ed.). São Paulo.

Vygotsky, L. S. (1978). *Mind in Society: Development of Higher Psychological Processes.* (J.-S. V. Cole M., Ed.) Cambridge, Massachusetts, London, England: Harvard University Press. doi:https://doi.org/10.2307/j.ctvjf9vz4

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large Databases. *ACM SIGMOD Record*. doi:https://doi.org/10.1145/235968.233324