# Análise de agrupamento aplicada a textos para construção de perfis de trabalhos de conclusão de curso

Barbara Barbosa Claudino da Silva, Sarajane Marques Peres, Bianca Canezim Letti,
Lucas Fernandes Brunialti, Marcelo Fantinato, Clodoaldo Aparecido Morais Lima
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
São Paulo - SP - 03828-000
Email: {barbara.barbosa.silva, sarajane, bianca.letti, lucas.brunialti, m.fantinato, c.lima}@usp.br

***Resumo*—Como requisito para conclusão de um curso semipresencial de especialização (pós-graduação *lato sensu*) de formação de educadores, cerca de 700 alunos desenvolveram uma monografia com tema de livre escolha. O requisito curricular que determinou a condição de "tema de livre escolha" e a grande quantidade de alunos participantes do curso tornou o trabalho dos coordenadores de curso bastante complexo no que tange ao desejo de analisar os perfis das temáticas escolhidas. Assim, um procedimento automático se fez necessário para apoiar o processo de análise e para isso, os resumos das monografias foram submetidos a uma análise de agrupamento, a qual tornou viável o trabalho de descoberta de perfis dos temas abordados. Nessa análise de agrupamento os algoritmos Self-Organizing Maps (SOM) e K-means foram aplicados em sequência. Os resultados obtidos, e descritos neste artigo, indicam que tal procedimento suportou com sucesso um processo de análise semi-automática tanto para o estabelecimento dos perfis quanto para o alinhamento entre as expectativas da coordenação de curso e as escolhas de temas por parte dos alunos.**

***Keywords*—*Self-Organizing Maps, K-Means, Análise de agrupamento, Mineração de texto, Análise de perfil, Descoberta de conhecimento.***

## I. Introdução

A atual diversidade de recursos para disseminar informação, comunicação e interação entre pessoas coloca o processo de aprendizado em um nível diferente do que era o padrão há alguns anos. Isso tem promovido o desenvolvimento de práticas inovadoras de ensino que propõem uma reinvenção da educação. Nesse contexto, a formação continuada de educadores é um dos pilares para que as inovações cheguem ao sistema educacional, modernizando-o e garantindo-lhe qualidade. Cursos oferecidos nas modalidades a distância e semipresencial são cada vez mais presentes e, devido às facilidades provenientes da inserção da tecnologia no cotidiano das pessoas, tais modalidades são capazes de atingir um grande número de pessoas, gerando a necessidade preemente do uso de ferramentas computacionais capazes de apoiar o trabalho de produção e disseminação de conhecimento e de avaliação do processo de ensino e aprendizagem.

Este artigo apresenta os primeiros resultados de uma análise semi-automática de resumos de 712 monografias produzidas no contexto de um curso de especialização semipresencial na área de formação de educadores. Essa análise semi-automática foi realizada por meio da resolução da tarefa de agrupamento, e constituiu-se como uma forma de descoberta de conhecimento em conteúdo textual. Como o tema da monografia foi de livre escolha do aluno, essa análise teve o objetivo de descobrir grupos de monografias cujo conteúdo possibilitasse a construção dos perfis dos temas abordados pelos alunos. Tal descoberta visou ainda apoiar a coordenação de curso no ajuste de suas ações para alinhamento de expectativas.

Como solução técnica para a análise de agrupamento, aplicou-se a técnica *Self-Organizing Maps* e o algoritmo K-*means*. Os resumos das monografias foram analisados automaticamente e subconjuntos de monografias foram descobertos. Depois, listas de termos (ou seja, palavras) foram extraídas para representar cada grupo. Essas listas foram analisadas empiricamente pelos analistas técnicos (ou seja, os proponentes da solução técnica) e por um dos coordenadores do curso. Como resultado, obteve-se uma base temática referente às monografias que permitem validar o atendimento de expectativas.

O conteúdo deste artigo está organizado da seguinte maneira: a Seção II apresenta uma contextualização sobre o curso de especialização onde o conteúdo (ou seja, as monografias) foi produzido, sobre as técnicas de agrupamento, e sobre os trabalhos correlatos; a metodolodia aplicada para resolver a tarefa de agrupamento e para analisar os resultados é apresentada na Seção III; os resultados produzidos são descritos e analisados na Seção IV; e, finalmente, considerações finais são delineadas na Seção V.

## II. Contextualização

Esta seção é dedicada a introduzir o contexto de aplicação do trabalho descrito neste artigo, as técnicas de análise de dados aplicadas, e trabalhos correlatos.

### A. O curso de especialização

O curso de especialização, chamado Ética, Valores e Cidadania na Escola, foi ministrado de forma semipresencial, e direcionado à atualização de profissionais da área de educação provenientes de escolas públicas, privadas, ONGs e instituições

que atuam na área social [1]. O conceito "*learning by doing*" foi o principal pilar da estruturação curricular, pedagógica e computacional [2] do curso, de forma que os alunos tiveram a oportunidade de [3]: conhecer novos paradigmas educacionais experenciando ativamente projetos sobre temáticas de ética e de cidadania; empregar diferentes ferramentas pedagógicas e linguagens na produção do conhecimento; e trabalhar de forma colaborativa e cooperativa na busca de soluções de problemas da realidade escolar em que atuavam. Ainda, houve uma preocupação em desenvolver uma perspectiva interdisciplinar que permitisse que os alunos articulassem as temáticas principais do curso e temas transversais, de forma a responder aos problemas sociais que norteiam o atual sistema educacional.

Diferentes artefatos de produção de conhecimento, com diferentes formatos para expor conhecimento, foram desenvolvidos pelos alunos durante o curso, incluindo uma monografia final. Os problemas investigados não foram definidos pelos docentes do curso mas sim pelos alunos, alterando o tradicional eixo de processos de aprendizagem nesse tipo de curso para essa área. Apesar disso, a coordenação de curso possuía a expectativa de que o cotidiano das escolas e organizações de origem dos alunos fossem naturalmente mapeados pelos alunos para a temática principal do curso: ética, valores e cidadania nas escolas. A avaliação dessas monografias foi realizada por bancas de avaliadores *ad hoc*, a maioria com formação em nível de doutorado. A qualidade das monografias individualmente foi avaliada de forma satisfatória. Porém, considerando o alto número de alunos concluintes (aproximadamente 700), não foi possível que os avaliadores conhecessem todo o conjunto de monografias. Isso dificultou que a coordenação de curso pudesse conhecer detalhadamente o perfil das monografias elaboradas em função dos temas escolhidos pelos alunos. Por isso, a extração automática de conhecimento realizada por esse trabalho é importante para evidenciar os perfis.

### B. Self-Organizing Maps e K-means

*Self-Organizing Maps* (SOM) são um tipo de rede neural artificial, caracterizada por um algoritmo de aprendizado competitivo, colaborativo e não-supervisionado. Via seu algoritmo de aprendizado, o SOM associa entradas $I$ ($i \in V$) a saídas $J$ ($j \in A$). As entradas são unidades que recebem exemplares de um conjunto de dados que determina o espaço de entrada vetorial $V$ do SOM. Saídas são representações das entradas organizadas em um mapa topológico (ou seja, uma grade $N$-dimensional de neurônios), o qual determina o espaço de saída $A$ do SOM. Entradas e neurônios são codificados em vetores ($x$ e $w$, respectivamente) no espaço $M$-dimensional, sendo $M \gg N$, geralmente.

O algoritmo de aprendizado adapta os vetores $w$ para que, em algum nível de abstração, eles possam representar as entradas $x$. SOM cria um mapa topográfico das entradas fornecidas a ele. As posições vetoriais dos neurônios indicam características estatísticas do conjunto de dados analisado. As relações de vizinhança topológica estabelecidas no espaço de saída de SOM indicam as relações topológicas existentes entre os dados. Os passos básicos do treinamento para SOM são apresentados no Algoritmo 1. Para mais detalhes sobre esse algoritmo, consulte [4].

**Algoritmo 1:** Passos básicos do algoritmo de treinamento de SOM

***1. Inicialização***: escolha valores aleatórios (ou use conhecimento *a priori*) para inicializar os vetores de pesos (neurônios) $w_j(0)$;

***2. Amostragem***: escolha, aleatoriamente, um exemplar $x$ do espaço de entrada;

***3. Cálculo de similaridade***: encontre o neurônio $J$ mais similar a $x$, na iteração $s$, usando uma medida de distância tal que $J = argmin_j ||x(s) - w_j(s)||$, $j = 1, 2, ..., k$, em que $k$ é o número de neurônios;

***4. Aprendizado***: ajuste os vetores de pesos de todos os neurônios aplicando $w_j(s+1) = w_j(s) + \alpha(s) * h_{j(x(s))}(s) * (x(s) - w_j(s))$, em que $\alpha$ é uma taxa de aprendizado e $h$ é uma função de vizinhança topológica centrada no neurônio vencedor $J$;

***5. Laço***: repita os passos de 2 a 4 até que nenhuma mudança significativa seja aplicada aos pesos.

SOM é útil em tarefas de redução de dimensionalidade e quantização do espaço, considerando um conjunto de dados com espaço de imersão alto ($D_V$). Essa utilidade é evidenciada pelo seu alto poder de descrição de relações de distância e topológicas existentes entre dados. Essa descrição ocorre por meio de vetores em baixa dimensão ($D_A$), em que geralmente $D_A << D_V$.

Há duas medidas de qualidade comumente usadas para avaliar o resultado produzido por SOM: erro de quantização, que mede o quanto SOM foi capaz de aproximar as propriedades estatísticas (e de densidade) dos dados; e erro topológico, que mede o quanto SOM foi capaz de preservar a topologia existente nos dados, ou seja, se as relações de vizinhança originais estão representadas no mapeamento. Para mais informações sobre essas medidas, consulte [5].

O resultado de quantização do espaço obtido via SOM permite realizar uma análise de agrupamento de dados. Porém, um procedimento de pós-processamento do resultado de SOM precisa ser realizado para observar os grupos evidenciados pela quantização. Uma estratégia comumente usada é aplicar um algoritmo simples de agrupamento, como K-*means*, sobre os neurônios de SOM. Assim, são procurados grupos em um espaço mais simples do que o espaço original dos dados. Os passos básicos de K-*means* são apresentados no Algoritmo 2. Para mais detalhes sobre esse algoritmo, consulte [6].

### C. Trabalhos correlatos

Até onde foi possível verificar, não foi encontrado nenhum trabalho que aplicasse técnicas de análise automática de textos em conteúdo gerado por alunos, com o objetivo de identificar os perfis referentes à escolha dos temas trabalhados na geração de tal conteúdo.

No contexto educacional, a descoberta de conhecimento via análise de textos é mais comumente vista como uma forma de apoiar diretamente o processo de ensino e aprendizagem. Exemplos de iniciativas nessa linha extraem tipos de conhecimento a partir de textos e usam esse conhecimento para: apoiar a produção textual [7], [8]; avaliar a qualidade dos textos [9], [10]; e analisar como o aluno interage com o ambiente de estudo [11]. Em geral, o desenvolvimento dessas iniciativas

**Algoritmo 2:** Passos básicos de K-*means*

***1. Inicialização:*** escolha o número $K$ de grupos a serem encontrados. Escolha valores aleatórios (ou use conhecimento *a priori*) para inicializar os vetores protótipos (centros de grupos) $c_j(0)$, em que $j = 1, 2, ..., K$;

***2. Amostragem:*** escolha, aleatoriamente, um exemplar $x$ do conjunto de dados;

***3. Cálculo de similaridade:*** encontre o protótipo $K$ mais similar a $x$, na iteração $s$, usando uma medida de distância tal que $C = argmin_j||x(s) - c_j(s)||, j = 1, 2, ..., K$;

***4. Atualização dos protótipos:*** ajuste os vetores protótipos de todos os grupos de forma que eles representem o centro do grupo, considerando todos os $x$ que estão associados a ele na iteração $s$;

***5. Laço:*** repita de (2) a (4) até que nenhuma mudança significativa seja aplicada aos vetores protótipos.

ocorre baseado em ferramentas que analisam a frequência dos termos que aparecem nos textos e as relações existentes entre esses termos, e que possibilita a visualização gráfica dos resultados, ou baseado em análises estruturais das frases que compõem os textos.

As relações existentes entre termos pode, além de evidenciar conhecimento sobre um texto, trazer conhecimento que está embutido em um *corpus* textual. Nesse caso, além das relações entre os termos dentro do texto, as relações dos termos no conjunto do *corpus* são importantes, e frequentemente é possível gerar modelos de tópicos que descrevem grupos de textos dentro do *corpus*, como é o objetivo do trabalho apresentado neste artigo. Esse tipo de análise é comum, por exemplo, em sistemas que pretendem recomendar um texto para leitura [12], [13], [14], sendo que nesses casos costuma-se também considerar o perfil do aluno/usuário ao propor a recomendação.

Em relação a análise de agrupamento em dados do tipo texto, os seguintes trabalhos guiaram a composição da metodologia usada no estudo apresentado neste artigo: [15], que aplicaram técnicas de redução de dimensionalidade (*Singular Values Decomposition* e *Principal Component Analysis*) e então de agrupamento (K-*means* e *Support Vector Clustering*); [16], que aplicaram SOM para agrupamento; e [17] que aplicaram algoritmos baseados em fatorização de matrizes para análise de co-grupos com intuito de explorar grupos de palavras *versus* grupos de textos.

## III. Metodologia

Nesta seção, é apresentada a metodologia aplicada para conduzir os experimentos e as análises realizados neste estudo. Para isso, o conteúdo da seção está dividido em duas partes: a descrição do *corpus* de documentos submetido à análise automática; e a apresentação da estratégia aplicada para encontrar os grupos de textos para então analisar o conhecimento embutido nessa estrutura de grupos.

### A. Corpus

O *corpus* submetido à análise é composto pelos resumos de 712 monografias apresentadas como requisito para completar um curso de especialização. Originalmente, 817 arquivos foram extraídos do banco de dados da plataforma computacional do curso. Porém, parte desses arquivos foram excluídos do *corpus* por duas razões: (a) duplicação de arquivos no banco de dados; e (b) ausência da seção de resumo[1]. Os 712 resumos que compõem o *corpus* foram submetidos a um pré-processamento que contou com três etapas:

- **Eliminação de *stop-words***: exclusão de palavras que são frequentes mas não possuem poder discriminatório em uma linguagem para poder contribuir para identificar o conteúdo tratado no texto. Neste trabalho, foi usado um conjunto clássico de *stop-words* composto principamente por artigos, preposições, conjunções e advérbios, acrescido de um subconjunto de palavras identificadas como comuns especificamente para esse *corpus* (abordaremos, aborda, apresenta, apresentar, trabalho, conclusão, especialização, pesquisa, presente, objetivo, professor).
- ***Stemming* e remoção de termos pouco e muito frequentes**: redução de palavras a seus radicais, para que formas plurais, gêneros e conjugações verbais sejam anuladas. O processo de *stemming* aplicado neste trabalho está baseado na remoção de afixos e, portanto, é dependente da língua em uso – o português. O *stemmer* PTStemmer[2] foi usado neste estudo. Nessa etapa, optou-se por manter os acentos e por separar palavras unidas por hífen, por observar o conjunto de termos gerados dessa forma, que se mostrou mais consistente. Ao final do processo de *stemming*, obteve-se um conjunto de 5.861 termos, sendo que o termo mais frequente apareceu 2.192 vezes no *corpus* pré-processado. Para fins de uso nos experimentos descritos aqui, uma faixa de frequência foi adotada de forma que termos que apareciam pelo menos 6 vezes no *corpus* e no máximo 1.821 vezes foram usados.
- **Construção de matrizes de representação dos textos**: construção de matrizes de textos e termos (ou seja, $resumos \times termos$) nas quais são representadas: a existência de um termo em um texto – matriz binária; a relação entre a frequência dos termos nos textos (*tf*) e a quantidade de textos do *corpus* que contém o termo (frequência inversa - *idf*) – matriz *tf-idf*; e a mesma relação porém normalizada – matriz *tf-idf* normalizada. Para mais detalhes sobre a construção das matrizes de representação, veja [18].

### B. Descoberta e análise de grupos

Para este estudo, a SOM Toolbox [19] foi usada para implementar o SOM. Foram geradas 16 configurações de SOM, combinando as variações dos parâmetros[3]: inicialização de pesos (aleatória ou linear); algoritmo de treinamento (sequencial ou em lote); tipo da grade topológica (retangular ou hexagonal); tipo de vizinhança (bolha ou gaussiana); número

[1] Embora o trabalho fosse monográfico, parte dos alunos o apresentaram em formatos alternativos, nos quais não havia a previsão de uma seção de resumo.

[2] https://www.openhub.net/p/ptstemmer

[3] Os valores dos parâmetros seguem nomenclatura apresentada em Alhoniemi et al. (2000).

de neurônios (escolhido automaticamente pela ferramenta). Cada uma das 16 configurações foi aplicada sobre três matrizes (binária, *tf-idf* e *tf-idf* normalizada), gerando 48 mapas. Os erros de quantização e topológico foram calculados para todos os mapas gerados. Essas duas medidas apoiaram a escolha dos mapas que seriam analisados em mais detalhes. Deu-se prioridade para os mapas com os menores erros de quantização já que esse erro retrata o quão bem os neurônios conseguiram representar a distribuição dos dados. No caso de empates ou resultados próximos para o erro de quantização, o erro topológico foi analisado.

A análise mais detalhada do resultado produzido via SOM foi realizada pela aplicação de K-*means*, considerando $k = 10$, sobre os neurônios de SOM. O valor de $k$ foi arbitrariamente escolhido para que a próxima fase de análise (manual e empírica) fosse viabilizada. Valores maiores geram grupos com poucos resumos, enquanto valores menores geram grupos com muitos resumos. K-*means* foi aplicado para os três melhores mapas (cf. Seção IV), dentre os 48 gerados.

Os grupos gerados via K-*means* foram analisados em função do tipo de matriz de representação usada. Para a matriz binária, foram geradas listas de palavras que apareciam exclusivamente em cada grupo. Para as matrizes de frequência relativa, foram geradas listas das dez palavras mais frequentes em cada grupo. A análise empírica das listas de palavras permitiu atribuir um tema para cada grupo. Essa estratégia também foi usada para os resumos agrupados por neurônio de SOM. Nesse caso, como cada neurônio possui menos resumos associados em comparação com os grupos gerados via K-*means*, foi possível realizar também uma análise usando apenas a frequência do termo *tf* no resumo específico, e comparar os temas empiricamente associados ao grupo de resumos de um neurônio com os respectivos títulos das monografias, tornando a análise mais precisa.

## IV. Resultados

A Tabela I apresenta a configuração dos melhores mapas para cada matriz de representação. A primeira análise refere-se aos resultados da matriz de representação binária (mapa 1). Dada a natureza da representação, não é útil realizar uma interpretação sobre a simples presença de uma palavra em um resumo, sem um peso associado. Assim, uma possibilidade é analisar as palavras que aparecem exclusivamente em cada grupo, e suas respectivas frequências. Dez grupos foram descobertos usando essa técnica e palavras-chaves foram extraídas seguindo essa abordagem, as quais indicam temas abordados nos seguintes quatro grupos de resumos: "desenvolvimento humano", "contraturno/coordenação pedagógica", "enfermagem", e "católico/nordestino". Assim, foi possível observar que a estratégia de agrupamento com representação binária é capaz de encontrar palavras-chave *vs.* resumos que estão bem caracterizados dentro do *corpus*.

A segunda análise refere-se aos resultados do uso das matrizes de representação com frequências relativas (mapas 2 e 3), considerando como descritor de um grupo de resumos a lista das dez palavras mais frequentes nele. As listas foram submetidas à uma interpretação humana realizada por um analista técnico e por um coordenador de curso. Na Tabela II, são apresentados os dez grupos gerados a partir dos mapas 2 e 3, respectivamente, incluindo as seguintes informações: número de resumos associados a cada grupo, as palavras mais significativas das listas[4] e as temáticas associadas pelos analistas.

Tabela I. Melhores mapas gerados para cada tipo de matriz de representação

| Id | **Representação** | **Parâmetros do mapa** |
|---|---|---|
| Mapa 1 | binária | inicialização de pesos linear, algoritmo de treinamento em lote, grade hexagonal, função de vizinhança bolha |
| Mapa 2 | *tf-idf* | inicialização de pesos aleatória, algoritmo de treinamento em lote, grade hexagonal, função de vizinhança gaussiana |
| Mapa 3 | *tf-idf* normalizado | inicialização aleatória, algoritmo de treinamento em lote, grade hexagonal, função de vizinhança bolha |

A interpretação realizada pelo coordenador de curso foi "às cegas", ou seja, sem que ele tivesse informações sobre a metodologia usada para o levantamento automático das listas de palavras. Por outro lado, o coordenador de curso é parcialmente influenciado por conhecer as temáticas trabalhadas no currículo do curso, o que não ocorre com o analista técnico, que por sua vez é parcialmente influenciado por conhecer a estratégia usada pelos algoritmos de agrupamento. Além disso, para certos casos, o analista técnico não foi capaz de associar uma temática à lista de palavras. Os grupos foram criados de forma bem definida, visto a pequena intersecção de palavras entre os grupos. Para o mapa 3 (ou seja, o segundo conjunto de grupos na Tabela II, representado com a letra (b)), não houve repetição de palavras entre os grupos. Já em relação ao mapa 2 (ou seja, o primeiro conjunto de grupos na tabela, representados com a letra (a)), apenas as seguintes três palavras apareceram entre as mais frequentes em dois diferentes grupos: "brincadeira", "discurso" e "militante".

A terceira análise, também executada sobre resultados provenientes do uso de frequências relativas (mapas 2 e 3, Tabela I), considerou grupos mais restritos de resumos. Apenas os resumos associados aos neurônios mais representativos de SOM foram analisados (nesse caso, K-*means* não foi considerado). Os resultados são apresentados na Tabela III, com a análise de temáticas realizada apenas pelo coordenador de curso. Nesse caso, duas descrições para cada grupo foram fornecidas, as quais, embora compostas por listas de palavras diferentes (a primeira obtida pela frequência relativa e a segunda pela frequência simples), levaram à associação de temas semelhantes, ou iguais, em certos casos. Esse era um comportamento esperado já que se trata de um mesmo grupo, descrito de formas diferentes. O uso de duas formas diferentes para descrever os grupos pode conferir mais precisão e robustez às análises; assim, os grupos $1a$, $3a$ e $3b$ podem ser considerados mais bem formulados e descritos. Nas descrições de cada grupo, considerando cada tipo específico de descrição, há também poucas palavras repetidas. As palavras que aparecem na descrição de dois ou três grupos são: "cidadania", "educação", "formação", "*bullying*" e "gerações". Porém, des-

[4]Foram excluídas palavras comuns que se caracterizam como *stop-words*.

Tabela II. DESCRIÇÃO DE GRUPOS OBTIDOS COM SOM/K-MEANS/FREQUÊNCIAS RELATIVAS

| Id. | # | Palavras [temas associados pelo analista técnico / temas associados pelo coordenador do curso] |
|---|---|---|
| 1a | 174 | coordenação, pedagógico, cargo, atribuição, função, equipe, expectativa, dependência, grupo **[administração escolar / responsabilidade]** |
| 2a | 62 | dança, círculo, conselho, fracasso, grêmio, gravidez, idade, assembleia, leitor, estudantil **[suporte ao aluno / juventude]** |
| 3a | 48 | enfermagem, música, feliz, extensão, descontentamento, território, gênero, conduta, religião, física **[– / fronteiras]** |
| 4a | 42 | católico, surdo, inglês, prova, brincadeira, linguagem, recreio, língua **[linguagem / comunicação]** |
| 5a | 10 | militante, povo, estigma, serviço, discurso, brasil, instrução, LBD, estatuto **[legislação e direitos / política]** |
| 6a | 101 | humano, distância, plágio, apostila, indígena, computação, inclusão, consumo, brincadeira **[inclusão e educação à distância / participação]** |
| 7a | 74 | alfabetização, canal, virtude, comemoração, recuperação, escolha, matemática, paz, abuso, empreendedor **[desempenho e aprendizado / inclusão]** |
| 8a | 52 | nordestino, visita, PBL, habilidade, alto, superdotado, permanência, publicação, população, representação **[– / sujeito**] |
| 9a | 41 | ansiedade, contrato, cultura, militante, orientação, adolescência, direção, discurso **[adolescência / controle]** |
| 10a | 108 | desenvolvimento humano, alimentação, mito, herói, afeto, desempenho, linguístico, produções, menino **[condições de aprendizado / sucesso]** |
| 1b | 4 | literatura, comparação, condicional, configuração, economia, mito, qualificação, renovação, profissão, conceitual **[qualificação profissional / trabalho ]** |
| 2b | 75 | território, extensão, palestra, déficit, matemática, estigma, proposição, produções, contextualização **[problemas no ensino / pedagogia ]** |
| 3b | 215 | almejar, atribuições, funções, contraturno, edições, *funk*, férias, genitores **[administração escolar / futuro ]** |
| 4b | 6 | ideal, racismo, negro, estatuto, modelo, administração, disposto, afro, conceber, entrega **[afrodescendentes / marcadores sociais]** |
| 5b | 32 | abuso, PBL, autoimagem, virtude, assinalar, interlocução, conduta, acreditar, dignidade, verdade **[conduta / virtude]** |
| 6b | 24 | descontentamento, audiovisual, religião, escolha, colega, teatro, roda, coerência, internet, conversa **[recursos para ensino / participação]** |
| 7b | 19 | ateliê, conselho, publicação, recreio, dirigir, expressões, piagetiano, digital, verbal, democratização **[construção e comportamento / linguagens]** |
| 8b | 101 | canal, idade, injustiça, compreensão, socioeducação, paz, alfabetização, internacional, patrimônio, superdotado **[valores / formação ]** |
| 9b | 213 | católico, desenvolvimento humano, enfermagem, nordestino, comemoração, feliz, herói, data, alimentação, raiva **[condições emocionais / sensações]** |
| 10b | 23 | empreendedor, solidariedade, computação, música, adaptação, adoção, representação **[empreendedorismo solidário / envolvimento]** |

sas, as três primeiras poderiam ser consideradas *stop-words*, pois fazem parte do tema central do curso de especialização em questão.

Finalmente, a última análise permitiu refinar o conhecimento adquirido no experimento aqui descrito. Os títulos dos trabalhos associados a um dos grupos – 2*b* cf. Tabela III – e listas de palavras que descrevem esse grupo foram contrastados. Nesse caso, o coordenador de curso associou cada título a uma das duas listas. Os títulos dos trabalhos e as associações realizadas são apresentadas na Tabela IV. Ao combinar o resultado de análises diferentes, verifica-se que na visão do coordenador de curso, a presença de uma ou mais palavras-chave em um contexto não leva, necessariamente, à mesma interpretação que palavras isoladas podem levar.

## V. CONCLUSÕES

Neste artigo, foi apresentado um experimento sobre análise de agrupamento de textos produzidos por alunos de um curso de especialização. Os resultados foram interpretados para apoiar a identificação dos perfis dos temas abordados pelos alunos. O experimento contou com variações referentes à representação dos textos, configuração de técnicas computacionais e perspectivas de interpretação de resultados. Cada variação permitiu produzir análises diferentes porém consistentes. O contexto temático homogêneo em que se inserem os 712 resumos analisados inseriu uma dificuldade no processo de análise automática. É comum encontrar trabalhos na literatura que analisam *corpus* formados por documentos textuais que versam sobre assuntos de diferentes temas. Portanto, esse experimento se configura como uma contribuição para a área de descoberta de agrupamentos num contexto de análise de textos.

A análise automática de textos via análise de agrupamento foi positiva. Do ponto de vista técnico, embora existam melhorias possíveis, os resultados permitiram concluir que o conhecimento obtido é consistente. Do ponto de vista da coordenação de curso, foi possível construir uma nova fotografia sobre o curso, uma vez que essa experiência serviu como mecanismo novo de análise que possibilitou aferir com mais detalhes as escolhas temáticas feitas pelos alunos. Além disso, a coordenação de curso entende que o conhecimento trazido por essa experiência deverá compor o aporte de informação a ser usado, junto a outros elementos, para tomada de decisão visando ao alinhamento de expectativas.

Evoluções desse estudo estão sendo realizadas, incluindo a análise da informação topológica, expressa pelas relações de vizinhança de SOM, que pode indicar como os grupos e as temáticas se posicionam quanto a relações de similaridade. Além disso, uma análise de co-agrupamento está em curso, de forma que se espera gerar informações mais refinadas sobre o relacionamentos entre grupos de palavras e grupos de resumos.

## AGRADECIMENTOS

Os autores deste artigo agradecem: a colaboração do Prof. Dr. Marcos Garcia Neira, da Faculdade de Educação da Universidade de São Paulo, um dos coordenadores do curso de especialização *Ética, Valores e Cidadania da Escola*; o suporte recebido do Núcleo de Apoio à Pesquisa - Novas Arquiteturas Pedagógicas.

Tabela III. DESCRIÇÃO DE GRUPOS OBTIDOS COM SOM/FREQUÊNCIAS RELATIVAS

| Id. | # | | Termos [temas associados pelo coordenador de curso] |
|---|---|---|---|
| 1a | 13 | *tf-idf* | empreendedor, moral, formal, crise, social, agir, contribuição, geração, proporcionar, valor **[ética]** |
| | | *tf* | educação, moral, social, valor, contribuição, formação, cidadania, formal, construção **[ética]** |
| 2a | 15 | *tf-idf* | extensão, território, técnica, plano, ingresso, competência, alienação, didática, região, livre **[espaço]** |
| | | *tf* | educação, aula, técnica, cidadania, desenvolvimento, formação, novo, plano, valor **[pedagogia]** |
| 3a | 17 | *tf-idf* | visita, hábil, alto, superdotado, permanência, vínculo, leitura, informal, currículo, interlocução **[formação]** |
| | | *tf* | educação, hábil, alto, visita, permanência, currículo, ensino, espaço, formação, leitura **[formação]** |
| 1b | 26 | *tf-idf* | contraturno, município, mitologia, repertório, cançoes, promotor, lógica, grego, previnir **[responsabilidade]** |
| | | *tf* | cultura, educação, ambiente, brincadeira, criança, currículo, saúde, cidadania, didática **[escola]** |
| 2b | 13 | *tf-idf* | licenciatura, psicologia, detrimento, reprodução, vítima, relativo, colaboração, gerações, disciplina **[pedagogia]** |
| | | *tf* | valor, disciplina, *bullying*, educação, inclusão, problema, conteúdo, estudo, família **[moral]** |
| 3b | 18 | *tf-idf* | psíquico, dicotomia, afastamento, desenho, mal, gerações, superior, estatuto, adolescência **[ruptura]** |
| | | *tf* | adolescência, ambos, *bullying*, conflituosa, criatividade, design, gestão, formação, demonstra **[problemas]** |

Tabela IV. TÍTULOS DOS TRABALHOS DO GRUPO 2*b* (TABELA III) *vs.* DESCRITORES DE GRUPOS

| Id. | Título do trabalho [lista de palavras associada pelo coordenador de curso] |
|---|---|
| 1 | Inclusão escolar: impedimentos e a concepção docente **[*tf*]** |
| 2 | Educação inclusiva no ensino fundamental das escolas públicas: percepção dos docentes **[*tf*]** |
| 3 | As constribuições da psicologia nos cursos de licenciatura: um olhar sobre a compreensão de licenciandos e licenciados sobre a disciplina Psicologia da Educação e sua influência na prática pedagógica **[*tf*]** |
| 4 | As influências da educação em valores para o desenvolvimento infantil **[*tf*]** |
| 5 | Educação e reprodutivismo: a perspectiva educacional e sua influência na formação do sujeito de valores **[*tf-idf*]** |
| 6 | Educação e valores no ambiente escolar: o caminho para a paz nas escolas e das escolas para a vida **[*tf*]** |
| 7 | A escola aos olhos do aluno indisciplinado **[*tf-idf*]** |
| 8 | A contribuição do processo de formação de um grupo teatral no desenvolvimento da convivência democrática na escola **[*tf-idf*]** |
| 9 | As percepções de alunos do ensino médio de uma escola pública estadual do município de Bauru - SP acerca do bullying: caracterização e reflexoes sobre sua menifestação no ambiente escolar **[*tf*]** |
| 10 | Bullying: e o professor com isso? Uma análise do papel do professor perante esse fenômeno em escolas do município de Agudos - SP **[*tf-idf*]** |
| 11 | Família e escola: a construção de valores **[*tf*]** |
| 12 | Gestão democrática na construção de valores na educação básica **[*tf-idf*]** |
| 13 | A construção de valores no ensino fundamental I **[*tf-idf*]** |

## REFERÊNCIAS


[1] U. F. Araújo, "A quarta revolução educacional: A mudança de tempos, espaços e relações na escola a partir do uso de tecnologias e da inclusão social," *ETD – Educ. Temática Digital*, pp. 31–48, 2011.

[2] S. M. Peres, M. Fantinato, U. F. Araújo, M. P. Soledade Junior, R. S. Freitas, F. M. Azevedo, and H. M. A. Barbosa, "Apoio ao aprendizado semipresencial: Uma aplicação no curso de especialização Ética, valores e saúde na escola," in *Anais do XXIII Simp. Bras. de Inform. na Educ.*, 2012, pp. 8–1–8–10.

[3] U. F. Araújo, M. C. Garbin, J. Franzi, V. A. Arantes, and C. C. O. Silva, "O uso de tecnologias educaionais na formação de professores para conteúdos de ética e cidadania: o curso de especialização semipresentação em Ética, valores e cidadania na escola," *Int. Studies on Law and Educ.*, vol. 19, pp. 37–46, 2015.

[4] S. Haykin, *Neural networks and learning machines*, 3rd ed. Prentice Hall, 2008.

[5] E. C. Kitani, "Mapeamento e visualização de dados em alta dimensão com mapas auto-organizados," Ph.D. dissertation, Escola Politécnica, Univ. de São Paulo, 2013.

[6] T. Rocha, S. M. Peres, H. H. Bíscaro, R. C. B. Madeo, and C. Boscarioli, "Tutorial sobre fuzzy-c-means e fuzzy learning vector quantization: Abordagens híbridas para tarefas de agrupamento e classificação," *Rev. de Inform. Teórica e Aplicada*, vol. 19, no. 1, 2012.

[7] M. Klemann, E. Reategui, and A. Lorenzatti, "O emprego da ferramenta de mineração de textos SOBEK como apoio à produção textual," in *Anais do XX Simp. Bras. de Inform. na Educ.*, 2009.

[8] M. Klemann, E. Reategui, and C. Rapkiewicz, "Análise de ferramentas de mineração de textos para apoio à produção textual," in *Anais do XXII Simp. Bras. de Inform. na Educ.*, 2011, pp. 1100–1103.

[9] J. C. S. Nobre and S. R. M. Pellegrino, "ANAC: un analisador automático de coesão textual em redação," in *Anais do XXI Simp. Bras. de Inform. na Educ.*, 2010.

[10] ——, "Avaliador automático de coesão textual em redação dissertativa - AVAC," in *Anais do XXI Simp. Bras. de Inform. na Educ.*, 2010.

[11] B. R. T. Azevedo, P. A. Behar, and E. B. Reategui, "Análise das mensagens de fóruns de discussão através de um software para mineração de textos," in *Anais do XXII Simp. Bras. de Inform. na Educ.*, 2011, pp. 20–29.

[12] T. Yoneya and H. Mamitsuka, "Pure: A PUBMED article recommendation system based on content-based filtering," *Genome Inform.*, vol. 18, pp. 267–276, 2007.

[13] C. Haruechaiyasak and C. Damrongrat, "Article recommendation based on a topic model for wikipedia selection for schools," in *Proc. of 11th Int. Conf. on Digital Libraries: Universal and Ubiquitous Access to Inf. – LNCS 5362*, G. Buchanan, M. Masoodian, and S. J. Cunningham, Eds., 2008, pp. 339–342.

[14] T. Tang, P. Winoto, and G. McCalla, "Further thoughts on context-aware paper recommendations for education," in *Recommender Syst. for Techn. Enhanced Learn.*, N. Manouselis, H. Drachsler, K. Verbert, and O. C. Santos, Eds., 2014, pp. 159–173.

[15] S. Jun, S. Park, and D. Jang, "Document clustering method using dimension reduction and support vector clustering to overcome sparseness," *J. Expert Syst. with Applications*, vol. 41, pp. 3204–3212, 2013.

[16] T. Kohonen, S. Kashi, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," *IEEE Trans. on Neural Networks*, vol. 11, no. 3, pp. 574–585, 2000.

[17] J. Yoo and S. Choi, "Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds," *Inform. Processing & Manag.*, vol. 46, no. 5, pp. 559–570, 2010.

[18] F. Sebastiani, "Text categorization," in *Text Mining and its Applications*. WIT Press, 2005, pp. 109–129.

[19] E. Alhoniemi, J. Himberg, J. Parhankangas, and Vesanto, "Som toolbox," http://www.cis.hut.fi/projects/somtoolbox/, 2000.