ANÁLISE DE AGRUPAMENTOS SOBRE TEXTOS: UM ESTUDO DOS RESUMOS DO BANCO DE TESES E DISSERTAÇÕES DA CAPES

Alexandre Leopoldo Gonçalves[1], Fernando Melo Faraco[2], João Artur de Souza[3], José Leomar Todesco[4], Ronnie Carlos Tavares Nunes[5]

Abstract: *The process of knowledge discovery in large volumes of information has a wide field of application. The main tasks of classification, clustering and association have been used in different areas of knowledge to make it possible to identify useful knowledge in large volumes of data. In this article, the application of data mining techniques, especially the K-Means clustering algorithm, is analyzed with the objective of verifying its effectiveness for the analysis of data from the Brazilian Open Data Portal, a public data repository organized and made available for the population. The dataset used for the application of the clustering algorithm was extracted from the information provided on the thesis and dissertation database made available by CAPES (Coordination of Improvement of Higher Education Personnel). The data were processed and inserted in the Apache Solr® platform where they were indexed, and the clusters were generated from the Carrot2 software, using the K-Means algorithm with customized configurations. The clusters were generated year by year and consolidated, with different configurations of the algorithm, making it possible to compare the obtained terms. It was concluded that the results of the used tools are directly related to the choice of the number of initial clusters, but the potential for discovering non-obvious clusters is obvious.*

Keywords: *Document Clustering; Open Data; Data Mining; K-Means; Knowledge Discover in Text.*

Resumo: O processo de descoberta de conhecimento em grandes volumes de informação tem um amplo campo de aplicação. As principais tarefas de classificação, agrupamento e associação têm sido utilizadas em diferentes áreas do conhecimento para tornar possível a identificação de conhecimento útil em grandes volumes de dados. Neste artigo, é analisada a aplicação de técnicas de mineração de dados, notadamente o algoritmo de agrupamento K-Means, com o objetivo de verificar sua efetividade para análise de dados oriundos do Portal Brasileiro de Dados Abertos, um repositório de dados público organizado e disponibilizado à população. O conjunto de dados utilizado para a aplicação do algoritmo de agrupamento foi extraído das informações disponibilizadas sobre o banco de teses e dissertações disponibilizadas pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Os dados foram tratados e inseridos na plataforma Apache Solr® onde foram indexados, sendo os agrupamentos gerados a partir do software Carrot^2, utilizando-se o algoritmo K-Means com configurações customizadas. Os agrupamentos foram gerados ano a ano e de forma consolidada, com diferentes configurações do algoritmo, tornando possível a comparação entre os termos obtidos. Concluiu-se que os resultados das ferramentas utilizadas estão diretamente relacionados com a escolha do número de agrupamentos iniciais, mas a potencialidade para a descoberta de agrupamentos não óbvios é evidente.

Palavras-chave*: Agrupamento de Documentos; Dados Abertos; Mineração de Dados; K-Means; Descoberta de Conhecimento em Texto.*

[1] Departamento de Engenharia do Conhecimento – Universidade Federal de Santa Catarina (UFSC). Florianópolis, SC - Brasil. E-mail: a.l.goncalves@ufsc.br

[2] Mestrando em Engenharia do Conhecimento – Universidade Federal de Santa Catarina (UFSC). Florianópolis, SC - Brasil. E-mail: farakeys@gmail.com

[3] Departamento de Engenharia do Conhecimento – Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC – Brasil. E-mail: jartur@gmail.com

[4] Departamento de Engenharia do Conhecimento – Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC – Brasil. E-mail: tite@egc.ufsc.br

[5] Mestrando em Engenharia do Conhecimento – Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC – Brasil. E-mail: rocatan@gmail.com

1 INTRODUÇÃO

O desenvolvimento da Internet, aliado às tecnologias de armazenamento e comunicação, propiciou ao governo brasileiro o desenvolvimento e a manutenção de diversas bases de dados, contendo registros das mais variadas áreas governamentais. Entretanto, o acesso a esses dados era um problema recorrente, pois não estavam disponíveis ao cidadão comum, com prejuízo inclusive de garantias constitucionais, como o direito de acesso à informação.

Segundo Milic, Veljkovic, & Stoimenov (2018), o termo "dados abertos" no domínio de governo eletrônico está diretamente relacionado com a transparência do trabalho no setor público. É a combinação de várias formatos e tipos de base de dados disponíveis publicamente através de repositórios governamentais (Kassen, 2018, p. 209).

Para formalizar o compromisso do país em relação aos dados abertos, o governo brasileiro, como membro co-fundador da Parceria para Governo Aberto (*Open Government Partnership* ou OGP), desenvolveu o Portal Brasileiro de Dados Abertos, hospedado no sítio http://www.dados.gov.br, formalizado no primeiro plano de ação de governo aberto, lançado na OGP e referenciado pelo Decreto sem número de 15 de setembro de 2011. O portal funciona como um catálogo que facilita a busca e utilização de dados publicados pelo governo, apesar de ainda não conter todas as informações de todos os órgãos.

Um dos *datasets* (conjunto de dados) disponibilizados no portal é aquele que traz os dados do banco de teses e dissertações da CAPES. Este conjunto de dados não traz as teses completas, mas um conjunto de metadados sobre teses e dissertações publicadas entre 1987 e 2012, dentro dos quais se encontram dados como a data da publicação, a instituição de origem, o título, a região, a área de concentração e o resumo da tese ou dissertação, configurando-se em um catálogo bastante completo sobre estes documentos. Para esta pesquisa foram selecionados os *datasets* correspondentes aos anos de 2008 a 2012 (período de 5 anos), totalizando 264.428 registros.

A análise exploratória de um conjunto de dados deste porte necessitaria de um grande esforço para extração de informação útil e padrões recorrentes, que não poderiam ser evidenciados com uma análise superficial por amostragem. A realização de simples consultas estruturadas (utilizando *Structured Query Language* - SQL, por exemplo) também não se mostra adequada para descoberta de padrões ou para a obtenção de *insights* sobre o conjunto de dados, uma vez que tais consultas são elaboradas com o objetivo de recuperação de

informações específicas. Este cenário é um campo propício para aplicação das técnicas de mineração de dados textuais.

Mineração de dados de uma forma geral é o processo de analisar grandes repositórios de informação no intuito de descobrir padrões e informações relevantes implícitas (Sumathi & Sivanandam, 2006). Dentre as operações possíveis, temos a associação, a classificação, regressão e o agrupamento (KAO et al., 2003). É um dos mais importantes paradigmas para análise avançada de negócios e suporte à decisão (Amani & Fadlalla, 2017). Com estas questões em foco surge a pergunta de pesquisa balizadora deste artigo: como a análise de agrupamentos pode ser aplicada a um conjunto de documentos obtidos a partir do banco de teses e dissertações da CAPES, de forma a identificar padrões relevantes?

O artigo foi estruturado da seguinte forma: Na seção 1 é realizada a apresentação do problema e a contextualização do mesmo. Na seção 2, é realizada uma breve revisão dos conceitos que suportam este artigo. Na seção 3, é apresentada a metodologia utilizada para construção desta pesquisa e como foi realizada aplicação do algoritmo de agrupamento nos dados selecionados, para obtenção dos grupos analisados. Na seção 4 é realizada uma discussão sobre esses resultados, apresentando as vantagens e alguns desafios do método utilizado e, por fim, são realizadas algumas considerações a respeito de trabalhos futuros.

2 MINERAÇÃO DE TEXTOS

Com o aumento expressivo da quantidade de informações coletadas através de transações e sensores, surge na década de 1990 o termo *Big Data*. Segundo Xindong Wu, Xingquan Zhu, Gong-Qing Wu e Wei Ding (2014), o *Big Data* representa grandes conjuntos de dados complexos obtidos através de diversas fontes autônomas, proporcionado pelo rápido desenvolvimento da capacidade de coleta e armazenamento dos dados.

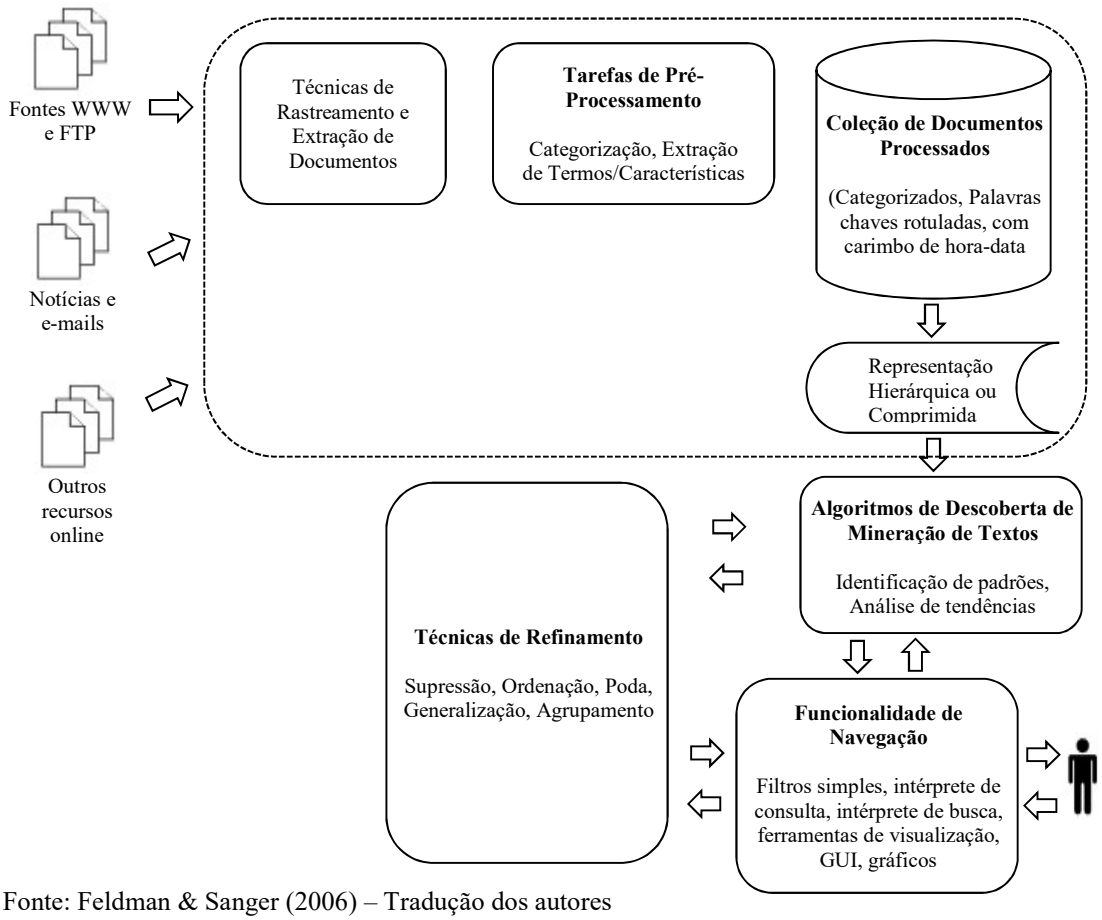
Neste cenário, se faz necessário o uso de técnicas adequadas para propiciar uma forma de descoberta de informações úteis e relevantes. Surgem os conceitos de Descoberta de Conhecimento em Textos (KDT) e Mineração de Textos (do inglês *Text Mining* - TM).

KDT é um processo não trivial de identificação de padrões implícitos, a partir de dados textuais, válidos, potencialmente úteis e compreensíveis (Figura 1). O termo “processo” implica que é composto de vários passos repetidos em múltiplas iterações, como preparação de dados, busca por padrões, avaliação do conhecimento e refinamento.

Um dos seus principais passos é a Mineração de Texto (TM), que consiste na

aplicação de algoritmos com o propósito de identificar padrões. É processo semiautomático de extração de conhecimento de uma grande quantidade de dados não estruturados, utilizando a combinação de técnicas de *data mining*, *machine learning*, processamento de linguagem natural, recuperação da informação e gerenciamento de conhecimento de grandes bases textuais (Delen & Crossland, 2008; Yang, Kleissl, Gueymard, Pedro & Coimbra, 2018).

Figura 1 – Arquitetura de um sistema genérico de mineração de texto



Fonte: Feldman & Sanger (2006) – Tradução dos autores

Entre as técnicas e algoritmos aplicáveis neste cenário, pode-se citar redes neurais artificiais, raciocínio baseado em caso, algoritmos genéticos, árvores de decisão, regras de associação, máquinas de vetores de suporte, regressão, mapas auto organizados, *k-nearest neighbor*, *naive bayes* e análise difusa (Amani & Fadlalla, 2017).

No que tange as tarefas de mineração de dados uma das técnicas mais utilizadas para exploração e análise de conjuntos de dados quando não se considera nenhuma classificação prévia, é a análise de agrupamento.

A organização de dados em grupos semelhantes é o modo mais fundamental para compreensão e aprendizagem (Jain, 2010) e o algoritmo *K-Means* é o método mais utilizado

quando se trata de grande número de dados numéricos dimensionais. Apesar do algoritmo *K-Means* ser simples e eficiente, ele tem algumas desvantagens (Yu, Chu, Wang, Chan & Chang, 2017), sendo: a) O número de *cluster* deve ser predeterminado, sendo difícil determinar o número apropriado; e) A determinação dos centroides iniciais irá afetar os *clusters* resultantes.

Com a finalidade de se diminuir essas desvantagens, vários trabalhos na literatura se propõem a aperfeiçoar o resultado do algoritmo. Gan e Ng (2017) pesquisaram uma implementação do *K-Means* para fornecer o agrupamento de dados e detecção de *outliers* simultaneamente, através da introdução de um agrupamento adicional que agrupa todos os *outliers*. (Mothukuri et al., 2017) propõem um método inicial de seleção de "sementes" eficiente, para melhorar o desempenho do método de filtragem do algoritmo *K-Means*, localizando os pontos de semente em áreas densas do conjunto de dados e bem separados.

Panapakidis e Christoforidis (2017) também propõem duas versões modificadas do algoritmo para melhorar a acurácia do agrupamento, uma vez que o *K-Means* é bastante sensível à seleção inicial dos centróides, enquanto Huang et al., (2016) propõem um novo tipo de algoritmo *K-Means* subespacial denominado *Time-Series K-Means* (*TSkmeans*) para dados de séries temporais de agrupamento. Há ainda abordagens que procuram incorporar o uso novas métricas de distância baseadas em divergência (Chakraborty & Das, 2017).

3 MÉTODOS DE PESQUISA

Neste artigo, é utilizado como algoritmo base o *K-Means* original. A quantidade de agrupamentos iniciais foi testada até se chegar a um número adequado, realizando-se uma análise qualitativa dos resultados obtidos.

A pesquisa de que trata este artigo foi dividida em 4 etapas: a) busca das definições basilares na literatura; b) seleção dos conjuntos de dados para aplicação do algoritmo de agrupamento; c) execução do processo de mineração de textos através da tarefa de agrupamento; e d) análise dos resultados.

Na primeira etapa, as definições de mineração de textos e do algoritmo de agrupamento selecionado para aplicação na pesquisa foram apresentadas através de uma pesquisa bibliográfica, permitindo o contato direto com as temáticas, a partir de material publicado em livros e artigos (Marconi & Lakatos, 2003).

Na segunda etapa, foi selecionada a base de dados para aplicação da pesquisa

propriamente dita. A base de dados foi escolhida utilizando-se de critérios como disponibilidade, confiabilidade, formato e relevância dos dados para realização do processo de agrupamento. A partir destes critérios, foram realizadas pesquisas no Portal Brasileiro de Dados Abertos, que contém diversos conjuntos de dados (*datasets*), dentre os quais foram selecionados os *datasets* disponibilizados pela CAPES, contendo dados e metadados obtidos a partir do banco de teses e dissertações da Instituição, entre os anos de 2008 e 2012. O período selecionado compreende os 5 (cinco) *datasets* mais recentes publicados até o momento (2017). Estes conjuntos de dados contêm dados e metadados de 264.428 teses ou dissertações, disponibilizados em formato de texto separado por vírgula (CSV) como dados abertos, possuindo volume e formato adequado para aplicação da pesquisa desejada.

A partir da seleção e análise dos conjuntos de dados, deu-se início à terceira etapa, na qual foi realizada a importação dos dados na plataforma de busca *Apache Solr*® para posterior agrupamento no software *Carrot2*. Este processo é detalhado na sequência.

O processo de agrupamento de dados de texto (*Document Clustering*) requer uma série de etapas adicionais para obtenção de um resultado satisfatório. Como na maioria dos processos de *data mining*, os dados precisam ser tratados, limpos, trabalhados, transformados, indexados e trazidos a uma base comum (utilizando-se, por exemplo, a normalização *tf-idf - frequency–inverse document frequency*) para que finalmente seja aplicado o algoritmo de *K-Means* sobre este resultado. É um processo trabalhoso, envolvendo várias etapas críticas.

Inicialmente os autores haviam optado pela realização de todo o processo utilizando-se a linguagem *Python*®, o que se mostrou adequado, mas pouco eficiente por uma série de motivos, entre os quais a necessidade de visualização gráfica de forma dinâmica dos agrupamentos obtidos. Esta funcionalidade precisaria ser construída, sendo necessário a construção de uma ferramenta de *software* específica, o que não é foco desta pesquisa. Optou-se, então, pela utilização de ferramentas já consolidadas, que implementassem todas as etapas do processo: a plataforma de indexação e busca *Apache Solr*® e o software *Carrot*^{2}® para realização dos agrupamentos.

O processo consistiu no tratamento dos dados quanto à sua codificação (WINDOWS-1252 para UTF-8) para posterior inclusão na plataforma *Solr*®. Cada *dataset* anual foi adicionado individualmente a base de documentos do *Solr*®, sendo criado um índice incremental para identificação de cada registro de tese ou dissertação. Os dados foram importados na sua íntegra, contendo os seguintes campos de dados: AnoBase, IdPrograma, Regiao, Uf, IdTese, SiglaIes, NomeIes, NomePrograma, GrandeAreaCodigo,

GrandeAreaDescricao, AreaConhecimentoCodigo, AreaConhecimento, AreaAvaliacao, DocumentoDiscente, Autor, EmailAutor, TituloTese, Nivel, DataDefesa, PalavrasChave, Volume, NumeroPaginas, BibliotecaDepositaria, Idioma, ResumoTese, IdLinhaPesquisa, URLTextoCompleto, LinhaPesquisa.

Uma vez criado o *core* (similar a uma base de dados) de pesquisa dentro do *Solr*®, foram realizadas consultas no campo "ResumoTese" para identificação da quantidade de documentos do Banco de Teses e Dissertações da Capes que continham determinadas palavras-chave em seu corpus (Tabela 1), considerando os anos de 2008 a 2012.

Tabela 1- Banco de Teses e Dissertações da CAPES

Ano	Documentos	"Clustering" OR Agrupamento"	"Interdisciplinar"	"Transdisciplinar"	"Engenharia" AND "Conhecimento"	"Engenharia do Conhecimento"
2008	46.750	345	330	58	88	2
2009	50.167	326	371	86	107	6
2010	50.903	344	367	77	118	4
2011	55.554	373	409	69	120	10
2012	61.054	461	441	83	144	6
Total	264.428	1.849	1.918	373	577	28

Fonte: Os autores (2018).

A partir das consultas realizadas, identificou-se que a utilização dos termos "Clustering" OR "Agrupamento" e "Interdisciplinar" trouxeram o maior número de resultados, e possuíam um maior potencial para execução do algoritmo de *Document Clustering*, sendo escolhido neste caso a utilização dos termos "*Clustering*" OR "Agrupamento" por se tratar do assunto deste artigo. Os termos "Transdisciplinar", "Engenharia" AND "Conhecimento" e "Engenharia do Conhecimento" apresentaram uma quantidade menor de documentos na base.

Definida a consulta para seleção dos documentos a serem agrupados, foi necessário realizar uma série de outros processos sobre estes documentos: a retirada das chamadas *stopwords* (palavras de parada, em livre tradução dos autores) que devem ser retiradas para não criar agrupamentos irrelevantes, a normalização dos termos utilizando *tf-idf* e, finalmente, a quarta etapa envolvendo a aplicação do algoritmo *K-Means* para construção dos agrupamentos. O processo de *stemming* (redução para o radical do termo), utilizado para diminuir a dimensionalidade do texto e trazer palavras similares a um radical comum, não foi aplicado explicitamente nesta pesquisa, sendo aplicado de forma não assistida e automática pela ferramenta *Carrot*^{2}®.

Após preparada a consulta com os termos selecionados ("*Clustering*" OR "Agrupamento"), iniciou-se a construção dos agrupamentos na ferramenta *Carrot*^{2}. Nesta pesquisa, priorizou-se a utilização do algoritmo *K-Means* para fins didáticos, mas estavam disponíveis na ferramenta também os algoritmos Lingo e STC (*Sufix Tree Clustering*). Outra possibilidade para construção dos agrupamentos seria incorporar à plataforma de indexação *Solr*® a própria ferramenta *Carrot*^{2}®, através de *plugin* específico. Neste caso, o resultado das consultas ao *Solr*® já viria agrupado, e a ferramenta *Carrot*^{2}® seria utilizado apenas para visualização.

Durante o processo de análise dos agrupamentos o *K-Means* foi utilizado com diferentes configurações para *k* (número de agrupamentos) e *n* (número de termos). A linguagem utilizada para as *stopwords* e para os agrupamentos foi o português, e o método de fatoração matricial foi o *partial single value decomposition* (*partial SVD*). Foram realizados testes com o método de fatoração do próprio *K-Means*, mas os resultados obtidos foram menos consistentes.

Após a geração dos agrupamentos, foi utilizada a ferramenta de visualização do *Carrot*^{2} denominada *FoamTree*, sendo também analisados os dados quantitativos e os termos (*labels*) identificados pelo algoritmo de agrupamento. Os resultados são apresentados na seção seguinte.

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Com os dados disponíveis na plataforma *Solr*®, foram montadas as consultas necessárias sobre os termos eleitos ("*Clustering*" OR "Agrupamento"), que já haviam se mostrado como excelentes candidatos para realização da pesquisa exploratória desejada na etapa anterior (tabela 1). A montagem das consultas foi realizada na ferramenta *Carrot*^{2}®, que permitiu a criação e análise dos agrupamentos pretendidos. Foram realizadas consultas para cada ano, e uma consulta consolidada contendo todos os anos (2008 a 2012).

Por padrão, o algoritmo *K-Means* vem configurado para 25 agrupamentos (k=25), e utiliza 3 termos para definição de cada grupo. O algoritmo de fatoração padrão é o *non-negative matrix factorization* e o número de iterações máximo para convergência do algoritmo é de 15. Esta configuração, apesar de registrar agrupamentos interessantes, não demonstrou ser adequada após análise dos documentos presentes em cada *cluster* (análise por amostragem).

Na Tabela 2, o algoritmo foi então reconfigurado para gerar k=15 agrupamentos, com

apenas *n*=1 termo para definição de cada grupo. O método de fatoração foi alterado para *partial singular value decomposition* (*SVD*).

Tabela 2 - Resultados do agrupamento para k=15, n=1, *partial* SVD

2008		2009		2010		2011		2012		2008-2012 (todos)	
Cluster	Docs	Cluster	Docs	Cluster	Docs	Cluster	Docs	Cluster	Docs	Cluster	Docs
Espécies	2	Espécies	8	Água	7	Genética	3	Sistema	0	Grupos	89
Sistema	0	Genes	8	Isolados	3	Pesquisa	0	Genética	6	Água	82
Documentos	9	Isolados	7	Área	3	Isolados	3	Espécies	3	Genética	60
Amostras	7	Genética	5	Espécies	0	Espécies	1	Saúde	3	Sistema	45
Genes	7	Desenvolvimento	4	Saúde	26	Genes	1	Genes	9	Algoritmo	143
Genótipos	7	Água	4	Algoritmo	5	Estrutura	5	Genótipos	4	Espécies	41
Populações	7	Algoritmo	3	Classificação	5	Floresta	23	Rede	2	Área	41
Isolados	5	Imagens	0	Acesso	2	Problemas	3	Algoritmo	0	Dados	16
Saúde	5	Acessos	9	Crianças	2	Área	3	Processo	9	Amostras	12
Algoritmo	4	Elementos	6	Modelo	2	Água	1	Algoritmo	6	Características	09
Modelo	2	Escola	6	Características	1	Crianças	0	Comunidade	5	Modelo	09
Valores	9	Gênero	4	Genes	7	Genótipos	8	Imagens	5	Populações	8
Molecular	8	Dados	2	Produção	7	Saúde	7	Pontos	4	Ambiente	2
Expressão	7	Atributos	0	Educação	8	Grupos	6	Concentração	2	Genes	0
Células	5	Modelo	9	Clones	5	Classificação	8	Unidades	2	Espécies	1
Total	344	Total	325	Total	343	Total	372	Total	460	Total	1848

Fonte: Os autores (2018).

Nesta configuração, percebe-se que há certa regularidade nos termos agrupados, assim como em sua distribuição (ocorrências), e que alguns termos aparecem consistentemente em diversos anos, assim como no agrupamento consolidado. Apesar da aparente regularidade encontrada, observa-se que alguns agrupamentos possuem potencial para serem mesclados (Genes e Genótipos, Genes e Genética, etc.), mas que não há expressividade suficiente no processo realizado para realizar este tipo de inferência somente a partir dos resultados dos agrupamentos, sendo necessária uma análise qualitativa do detalhamento de cada agrupamento para que sejam obtidas tais conclusões. Efetuando-se breve análise amostral destes agrupamentos, percebe-se que os termos "agrupamento" e "*clustering*" possuem ampla aplicação no campo da genética, que é o contexto de que tratam aqueles agrupamentos (genes, genótipo, genética, etc).

Na Tabela 3, apresenta-se os resultados com nova configuração do algoritmo para gerar *k*=15 agrupamentos, com dois termos desta vez (*n*=2), sendo possível verificar um nível maior de expressividade nos agrupamentos.

Tabela 3- Resultados do agrupamento para k=15, n=2, partial SVD.

2008		2009		2010		2011		2012		2008-2012 (todos)	
Cluster	Docs	Cluster	Docs	Cluster	Docs	Cluster	Docs	Cluster	Docs	Cluster	Docs
Genética, Acessos	33	Populações, Atributos	30	Genética, Genótipos	37	Acessos, Floresta	31	Ambientais, Sistema	50	Grupos, Genótipos	189
Genes, Expressão	31	Sistema, Fatores	29	Genes, Setor	29	Sistema, Rio	31	Genética, Acessos	46	Isolados, Água	182
Sistema, Serviços	31	Pontos, Nível	27	Água, Elementos	28	Genótipos, Alimentação	30	Espécies, Floresta	43	Genética, Acessos	160
Espécies, Gênero	29	Rede, Problema	27	Cultivares, Algoritmo	26	Área, Produção	28	Saúde, Óleo	43	Sistema, Maior	145
Cultivares, Comunidade	27	Espécies, Química	26	Técnicas, Sistema	26	Grupos, Teor	27	Isolados, Genes	39	Dados, Algoritmo	143
Algoritmo, Saúde	25	Genótipos, Frutos	26	Isolados, Proteína	23	Famílias, Saúde	26	Genótipos, Água	34	Espécies, Método	141
Amostras, Sequências	24	Genes, Escola	24	Atividades, Gestão	22	Pesquisa, Atividade	24	Rede, Qualidade	32	Espécies, Área	141
Teor, Valores	24	Isolados, Gênero	23	Grupos, Associação	22	Santa, Comunidade	24	Modelo, Algoritmo	30	Dados, Método	116
Genótipos, Óleo	19	Genética, Acessos	22	Espécies, Vegetação	21	Caracteres, Frutos	23	Processo, Aplicação	29	Dados, Amostras	112
Populações, Isolados	19	Fragmentos, Floresta	18	Populações, Educação	21	Classificação, Qualidade	23	Algoritmo, Identificação	26	Grupos, Características	109
Qualidade, Distribuição	18	Empresas, Milho	17	Modelos, Estimativas	20	Genética, Populações	22	Classificação, Imagens	25	Modelo, Método	109
Água, Anos	18	Imagens, Saúde	17	Espécies, Área	19	Rede, Proposta	22	Comunidade, Riqueza	25	Genética, Populações	88
Base, Desempenho	16	Algoritmo, Problema	14	Acessos, Floresta	18	Modelo, Pesquisa	21	Amostras, Pontos	14	Dados, Ambiente	82
Grupos, Desenvolvimento	15	Amostras, Água	13	Saúde, Animais	16	Algoritmo, Problemas	20	Concentração, Efeito	12	Grupos, Genes	80
Resistência, FLORESTA	15	Grupos, Elementos	12	Crianças, Variedades	15	Genes, Expressão	20	Dados, Unidades	12	Espécies, Maior	51
Total	344	Total	325	Total	343	Total	372	Total	460	Total	1848

Fonte: Os autores (2018).

Nesta configuração percebe-se que há uma maior expressividade nos grupos, e também algumas curiosidades, como agrupamentos entre palavras díspares: Algoritmo e Saúde (2008), Genes e Escola (2009), Caracteres e Frutos (2011) entre outros.

É possível verificar também que na consulta consolidada (2008-2012), alguns grupos provavelmente poderiam ser mesclados, ou até mesmo poderia ser gerado um agrupamento diferente para eles, como no caso dos agrupamentos (Dados, Algoritmo), (Dados, Método), (Dados, Amostras).

Conforme se aumenta o número de agrupamentos (*k*), passa a ser mais interessante a análise dos grupos utilizando-se das ferramentas de visualização disponibilizadas pelo *Carrot*^{2}®, que dispõe os grupos em um mapa, com tamanhos diferenciados de acordo com a quantidade de documentos.

Para o grupo consolidado (2008-2012), utilizando-se um *k*=50 e *n*=2, fatoração *partial SVD* (Figura 2), percebe-se a predominância de agrupamentos relacionados à genética, genes,

dados, grupos e outros, cuja visualização é bastante facilitada com o uso da ferramenta *FoamTree* do *Carrot*^{2}®. Estes agrupamentos são os que contêm o maior número de documentos (figuras maiores), permitindo supor que a maioria dos documentos obtidos nesta pesquisa ("*Clustering*" OR "Agrupamento") são da área das ciências da saúde ou ciências biológicas, e não da computação.

Figura 2 - Agrupamentos obtidos (2008-2012) utilizando-se k=50, n=2 e *partial* SVD.

Fonte: Os autores (2018)

5 CONSIDERAÇÕES FINAIS

A aplicação de análise de agrupamentos sobre conjuntos de documentos constitui-se em uma tarefa importante para se obter a essência e a composição dos dados, indo além das informações contidas em seus metadados ou sua classificação. Importante ressaltar que a aplicação de algoritmos de agrupamento é apenas uma das etapas de um processo bem mais amplo de mineração de dados ou textos, como já foi detalhado, e que deve ser acompanhado com bastante cuidado desde o seu início, pois cada uma das etapas pode afetar o resultado final deste processo.

A principal vantagem deste processo é a possibilidade de descoberta de agrupamentos não óbvios, obtidos a partir da análise de similaridade dos termos presentes nos documentos que fazem parte de determinado grupo. Para grandes conjuntos de dados, esta é uma tarefa difícil de realizar por meio de consultas estruturadas ou por análise amostral, e para isso, a utilização de algoritmos voltados ao agrupamento de dados se mostra adequado, desde que estes sejam configurados de maneira correta e o processo de preparação prévia dos dados seja

seguido.

Quanto aos resultados obtidos na aplicação do algoritmo, apesar de satisfatórios, é possível que com a utilização de diferentes configurações envolvendo, por exemplo, a remoção de *stopwords* e aplicação de *stemming*, a reconfiguração do número de agrupamentos (*k*), a definição do número de termos em cada agrupamento (*n*) e também do processo de fatoração matricial, fosse possível obter resultados de maior impacto.

Ainda que seja possível melhorar a qualidade dos agrupamentos para os termos pesquisados, a maior contribuição desta pesquisa consiste em evidenciar a potencialidade do método, como forma de obtenção de agrupamentos não óbvios em grandes bases de dados, onde a utilização de técnicas estatísticas não se mostra adequada. Visto que o procedimento é viável e reprodutível, pode ser aplicado em diversos outros conjuntos de dados, para diversas aplicações.

Como trabalhos futuros, pretende-se realizar uma análise mais detalhada e refinada em relação aos ajustes possíveis do algoritmo de agrupamento, como uma possível correlação do número de agrupamentos criados (*k*) e o contexto obtido nos agrupamentos, visando aprimorar a precisão do processo. Vislumbra-se ainda a aplicação das técnicas aqui descritas em banco de dados de teses e dissertações que disponibilizem o texto completo, obtendo como resultado prático não somente os agrupamentos, mas também a indexação e disponibilização de documentos para consulta por meio do *Solr*®. Existem ainda diversos outros conjuntos de dados que poderiam trazer *insights* interessantes, incluindo-se o suporte ao usuário (agrupamento dos principais tipos de chamado realizados por usuários), processo eletrônico judicial e administrativo (agrupamento de sentenças e de petições iniciais), sistemas de prontuários médicos, entre outros.

6 AGRADECIMENTOS

Agradecemos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo financiamento desta pesquisa.

REFERÊNCIAS

Amani, F. A., & Fadlalla, A. M. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*, *24*, 32–58.

Chakraborty, S., & Das, S. (2017). K−Means clustering with a new divergence-based distance

metric: Convergence and performance analysis. *Pattern Recognition Letters*, *100*, 67–73.

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, *34*(3), 1707–1720.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press.

Gan, G., & Ng, M. K. P. (2017). K-Means Clustering With Outlier Removal. *Pattern Recognition Letters*, *90*, 8–14.

Huang, X., Ye, Y., Xiong, L., Lau, R. Y. K., Jiang, N., & Wang, S. (2016). Time series k-means: A new k-means type smooth subspace clustering for time series data. *Information Sciences*, *367–368*, 1–13.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666.

Kassen, M. (2018). Open data and its institutional ecosystems : A comparative cross-jurisdictional analysis of open data platforms. *Canadian Public Administration*, *61*(1), 109–129.

Marconi, M. de A., & Lakatos, E. M. (2003). *Fundamentos De Metodologia Cientifica* (5ª Ed.). Editora Atlas S.A.

Milic, P., Veljkovic, N., & Stoimenov, L. (2018). Comparative analysis of metadata models on e-government open data platforms. *IEEE Transactions on Emerging Topics in Computing*, *6750*(c), 1–1.

Mothukuri, U. K., Reddy, B. V., Reddy, P. N., Gutti, S., Mandula, K., Parupalli, R., … Magesh, E. (2017). Improvisation of learning experience using Learning Analytics in eLearning.

Panapakidis, I. P., & Christoforidis, G. C. (2017). Implementation of modified versions of the K-means algorithm in power load curves profiling. *Sustainable Cities and Society*, *35*(July), 83–93.

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications*. *Studies in Computational Intelligence* (Vol. 29).

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, & Wei Ding. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107.

Yang, D., Kleissl, J., Gueymard, C. A., Pedro, H. T. C., & Coimbra, C. F. M. (2018). History and trends in solar irradiance and PV power forecasting: A preliminary assessment and

review using text mining. *Solar Energy*, (November), 0–1.

Yu, S.-S., Chu, S.-W., Wang, C.-M., Chan, Y.-K., & Chang, T.-C. (2017). Two Improved k-means Algorithms. *Applied Soft Computing*.