Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 25, Num. 04 (2018) 74-102

RESEARCH ARTICLE

# Use of text mining techniques for unsupervised organization of digital procedural acts

## Utilização de técnicas de mineração de texto para organização não supervisionada de atos processuais digitais

Alfredo Silveira Araújo Neto[1]*, Marcos Negreiros[2]

**Abstract:** The rapid advances in technologies related to the capture and storage of data in digital format have allowed to organizations the accumulation of a volume of information extremely high, constituted a higher proportion of data in unstructured format, represented by texts. However, it is noted that the retrieval of useful information from these large repositories has been a very challenging activity. In this context, data mining is presented as a self-discovery process that acts on large databases and enables the knowledge extraction from raw text documents. Among the many sources of textual documents are electronic diaries of justice, which are intended to make public officially all the acts of the Judiciary. Despite the publication in digital form has provided improvements represented by the removal of imperfections related to divulgation at printed format, it is observed that the application of data mining methods could render more rapid analysis of its contents. In this sense, this article establishes a tool capable of automatically grouping and categorizing digital procedural acts, based on the evaluation of text mining techniques applied to groups determination activity. In addition, the strategy of defining the descriptors of the groups, that is usually conducted based on the most frequent words in the documents, was evaluated and remodeled in order to use, instead of words, the most regularly identified concepts in the texts.

**Keywords:** Data mining — Heuristic — Combinatorial optimization — Bio-inspired computing

**Resumo:** Os rápidos avanços das tecnologias relacionadas à captura e ao armazenamento de dados em formato digital têm permitido às organizações o acúmulo de um volume de informações extremamente elevado, constituído em maior proporção por dados em formato não estruturado, representados por textos. Contudo, observa-se que a recuperação de informações úteis a partir desses grandes repositórios tem-se revelado uma atividade bastante desafiadora, que em geral não admite a utilização de técnicas tradicionais. Neste contexto, a mineração de dados apresenta-se como um processo de descoberta automática que age sobre grandes bancos de dados e que possibilita a extração de conhecimento a partir de documentos textuais brutos. Dentre as inúmeras fontes de documentos textuais encontram-se os diários de justiça eletrônicos, que têm como propósito tornar públicos de modo oficial todos os atos do Poder Judiciário. Não obstante a publicação em formato digital tenha proporcionado melhorias representadas pela supressão de imperfeições pertinentes à divulgação em formato impresso, verifica-se que a aplicação de métodos de mineração de dados poderia tornar mais célere a análise dos seus conteúdos. Neste sentido, este trabalho estabelece uma ferramenta apta a agrupar e categorizar de forma automática atos processuais digitais, a partir da avaliação de técnicas de mineração de textos aplicadas à atividade de determinação de grupos. Adicionalmente, a estratégia de definição dos descritores dos grupos, que em geral é conduzida com base nas palavras mais frequentes existentes nos documentos, foi avaliada e remodelada no intuito de empregar, ao invés das palavras, os conceitos mais regularmente identificados nos textos.

**Palavras-Chave:** Mineração de dados — Heurística — Otimização combinatória — Computação bioinspirada

[1] *Departamento de Computação, Universidade Estadual do Ceará, Brasil*
[2] *Departamento de Computação, Universidade Estadual do Ceará, Brasil*
*__Autor correspondente__:alfredosilveira@yahoo.com.br
**DOI:** https://doi.org/10.22456/2175-2745.83581 • **Received:** 05/06/2018 • **Accepted:** 07/09/2018

## 1. Introdução

Os rápidos avanços das tecnologias relacionadas à coleta e ao armazenamento de dados têm permitido às organizações o acúmulo de um volume de informações extremamente elevado [1]. Segundo estimativas, observa-se que no período compreendido entre 2006 e 2011, a quantidade de informações produzidas em formato digital passou de 180 exabytes (aproximadamente cento e oitenta bilhões de gigabytes) por ano para 1,8 zettabyte a cada ano, e que cerca de 80% desses dados estão em formato não estruturado, representados de forma significativa por textos [2, 1]. Entretanto, a recuperação de informações úteis a partir desses grandes repositórios tem-se revelado uma atividade bastante desafiadora, na qual a utilização de técnicas tradicionais muitas vezes não pode ser realizada em virtude da dimensão do conjunto de dados e de sua natureza ocasionalmente não trivial. Neste contexto, a mineração de dados se apresenta como um processo de descoberta automática, que age sobre grandes bancos de dados com o objetivo de identificar padrões úteis que, de outra maneira, permaneceriam desconhecidos [1]. A sua utilização possibilita a extração de conhecimento a partir de documentos textuais brutos, acrescentando informações extensivas sobre os mesmos, tais como agrupamentos de documentos similares, o que representam uma melhoria na recuperação de dados relevantes por parte das organizações [3].

Dentre as inúmeras fontes de documentos textuais, disponibilizadas em formato digital, encontra-se o Diário de Justiça Eletrônico. Este instrumento, estabelecido pela lei número 11.419/2006, permite a eliminação do registro impresso sem detrimento ao seu propósito de tornar público de modo oficial todos os atos do Poder Judiciário [4]. O seu formato promove a melhoria da publicidade realizada pelos diários de justiça, por intermédio da supressão de imperfeições pertinentes à divulgação em modelo impresso que, em algumas circunstâncias, podem dificultar o alcance de uma transparência efetiva, ao mesmo tempo em que facilita a recuperação de informações específicas pelo uso de diversas formas de consultas instantâneas, como por número de processo, nome do advogado, nome das partes, órgão jurisdicional competente etc. [5].

A despeito dos benefícios proporcionados pelos diários de justiça eletrônicos, verifica-se que a ausência da aplicação de métodos capazes de classificar os atos processuais que os compõem, torna a análise dos mesmos tão demorada quanto à exigida quando somente a divulgação em formato tradicional impresso encontrava-se disponível, pois os interessados, em particular os profissionais advogados, têm que invariavelmente ler o conteúdo das intimações publicadas a fim de determinar o seu nível de relevância. Neste aspecto, tarefas eficazes e eficientes de mineração de dados podem ser empregadas, com o objetivo de organizar a coleção de textos, nesta circunstância representada pelos atos processuais, em grupos de documentos com temas e assuntos semelhantes, aptos a fornecer uma descrição breve e representativa do conhecimento implícito nos textos, minorando o esforço e o tempo demandados na eventual avaliação de documentos de pouca importância [3].

Diante deste contexto, este trabalho teve como objetivo desenvolver um sistema de apoio à decisão (SAD) apto a classificar forma automática os atos processuais disponibilizados por meio dos diários de justiça eletrônicos, e, com este propósito, as seguintes condutas foram empreendidas: *i*) avaliar técnicas de mineração de textos utilizadas para prover o agrupamento e categorização de documentos; *ii*) incorporar ao SAD desenvolvido a técnica melhor avaliada; *iii*) empregar a ferramenta elaborada de modo integrado a um sistema de gestão de informações jurídicas no intuito de agrupar e categorizar automaticamente os atos processuais digitais.

A estruturação e a organização não supervisionada das coleções de publicações, com o emprego de métodos que não exigem conhecimento anterior acerca dos dados analisados, serão realizadas pela aplicação de um processo de mineração de textos constituído de três etapas principais: *i*) coleta de documentos; *ii*) pré-processamento; *iii*) extração de padrões com agrupamentos. Na fase de coleta de documentos, os atos processuais dos diários de justiça eletrônicos serão importados para o banco de dados do sistema de gestão de informações, por meio da utilização dos serviços de uma empresa especializada em captura de informações de diários oficiais digitais. Na etapa de pré-processamento, os textos escritos em linguagem natural serão submetidos a atividades de tratamento e padronização, seleção de termos mais significativos e representação em formato estruturado, passível de manipulação por algoritmos de agrupamento de textos, com a preservação das principais características dos dados. O tratamento e a padronização dos textos serão realizados por meio de procedimentos de lematização e remoção dos termos irrelevantes, conforme orientações descritas em [6]. A representação em formato estruturado será obtida com a utilização do modelo espaço vetorial, no qual cada ato processual será um elemento no espaço $m$-dimensional e cada dimensão representará uma característica da coleção de documentos com seu valor sendo dependente do grau de relacionamento entre a característica e o documento que a contém. Na fase de extração de padrões com agrupamentos, os objetos serão organizados em grupos. A atividade de determinação dos grupos será modelada como um problema de otimização ao qual serão aplicados os métodos iterativos pesquisa harmônica, algoritmo genético e *K-means*, com o emprego da distância do coseno como o índice de determinação da dissimilaridade entre os documentos. Os métodos iterativos serão confrontados entre si em termos da performance de execução e da qualidade da solução, computada por meio da aplicação de índices de validação de agrupamentos, a fim de que o algoritmo mais apropriado possa ser indicado. Adicionalmente e tendo em conta a inabilidade dos métodos iterativos analisados em reconhencer o número de grupos naturais presente em uma coleção de padrões, o algoritmo iterativo melhor avaliado será subsequentemente comparado ao algoritmo de passagem única $C^3M$, estabele-

cido por [7] e que possui a distintiva característica de, supostamente, determinar o número de grupos $K$ entre os quais os objetos da coleção de documentos devem ser distribuídos. Em sendo mal sucedido optar-se-á por avaliar outros métodos de identificação automática de grupos, a exemplo do CLUES [8] e do IGN [9], a fim de verificar se seu comportamento é condizente com o esperado, ressaltando-se que a despeito de serem mecanismos também capazes estabelecer uma partição sobre uma coleção de objetos, somente o aspecto relacionado à determinação do número de grupos será observado para o contexto deste estudo. Por fim, para o método de particionamento e de reconhecimento do número de grupos melhor avaliado, que deverá ser integrado ao sistema de gestão de informações jurídicas, haverá a incorporação de estratégias para seleção dos descritores dos grupos, a fim de que auxiliem a interpretação dos resultados na medida em que indiquem o significado dos agrupamentos obtidos aos interessados.

Além da seção 1, representada pela presente introdução, este trabalho compreende seis seções adicionais, as quais estão organizadas conforme especificado a seguir. A seção 2 apresenta alguns dos aspectos pertinentes à mineração de dados e à mineração de textos, considerados relevantes para o contexto deste trabalho. A seção 3 descreve a atividade de análise de agrupamentos, a seção 4 formaliza o problema do agrupamento de documentos, enquanto que a seção 5 refere algoritmos de particionamento que podem ser aplicados na sua resolução. A seção 6 descreve os experimentos de avaliação dos algoritmos de agrupamento mencionados na seção 5. A seção 7 estabelece um método de particionamento e categorização de atos processuais digitais, definido com base nos resultados apresentados na seção 6. Finalmente, a seção 8 apresenta as conclusões obtidas consoante os resultados dos experimentos.

## 2. Mineração de dados e mineração de textos

A mineração de dados é um processo de descoberta automática de conhecimento em grandes repositórios de dados. Corresponde a um conjunto de técnicas que atuam sobre grandes bancos de dados a fim de identificar padrões úteis, que de outra forma, permaneceriam desconhecidos. Em geral, as tarefas de descoberta de informação nem sempre podem ser consideradas mineração de dados. Atividades tais como a procura de registros específicos utilizando um sistema gerenciador de banco de dados ou o emprego de mecanismos de busca da Internet para localização de páginas *web* são, na verdade, tarefas relacionadas à área de recuperação da informação. Apesar de serem importantes e de muitas vezes utilizarem algoritmos e estruturas de dados sofisticadas, essas atividades são baseadas em técnicas tradicionais da ciência da computação e em recursos comuns dos sistemas gerenciadores de bancos de dados que proporcionam a organização e a recuperação eficiente das informações [1].

A mineração de textos consiste em um conjunto de tecnologias que têm como finalidade analisar e processar extensas coleções de documentos que estejam em um formato desestruturado ou semi-estruturado, e que têm como particularidade em comum a necessidade de converter o texto em um formato numérico estruturado de modo que algoritmos analíticos possam ser aplicados. Essencialmente, todas as técnicas de mineração de textos procuram endereçar a dificuldade relacionada a como examinar o exponencialmente crescente volume de dados textuais, a fim de obter dos mesmos informações relevantes, haja vista que enquanto a quantidade de textos originados em formato eletrônico eleva-se demasiadamente, observa-se que o volume de informações capaz de ser analisado por uma pessoa ou por um conjunto de indivíduos permanece inalterado. Se, por exemplo, alguém recebe diariamente até 10 cartas, individualmente constituídas por somente uma página, então será possível no mesmo dia ler todas as correspondências, refletir sobre o seu conteúdo, e posteriormente elaborar uma resposta. Entretanto, para um indvíduo que cotidianamente recebe entre 100 e 200 mensagens eletrônicas, uma conduta semelhante provavelmente não poderá ser admitida, e, certamente, estratégias que sejam aptas a processar esse expressivo volume de dados textuais devem ser desenvolvidas [10].

## 3. Análise de agrupamentos

Constituindo-se como um dos elementos que fazem parte do conjunto de tarefas descritivas da mineração de dados, a análise de agrupamentos pode ser definida como a organização de uma coleção de objetos em grupos baseada em uma medida de similaridade. De modo intuitivo, objetos que pertencem a um dado grupo são mais similares entre si do que quando comparados a quaisquer outros objetos que fazem parte de um grupo distinto, e, quanto maior a similaridade (ou homogeneidade) dentro de um grupo e maior a diferença entre os grupos, melhor ou mais distinto será o agrupamento [11, 1].

A análise de grupos é uma atividade humana importante. Desde o início da infância, as crianças aprendem a distinguir entre cães e gatos, ou entre animais e plantas, por meio de um processo contínuo de melhoria de esquemas semiconscientes de agrupamento. Métodos automatizados de agrupamento podem identificar regiões esparsas e densas no espaço de objetos, reconhecendo a distribuição integral dos elementos, além de correlações interessantes entre os atributos de dados. A análise de agrupamentos tem sido amplamente utilizada em muitas aplicações, a exemplo da pesquisa de mercados, do reconhecimento de padrões, da análise de dados ou do processamento de imagens. A análise de agrupamentos pode do mesmo modo, ajudar na determinação de áreas geográficas similares, por meio da observação das informações registradas em um banco de dados geográfico, além de facilitar a classficação de documentos oriundos da *web*, no intuito de promover a descoberta de informações relevantes implícitas nos conteúdos dos textos [12].

### 3.1 Categorização dos problemas de agrupamento

Um agrupamento é um tipo de classificação imposta a um conjunto finito de objetos, no qual o relacionamento entre os mesmos pode ser representado por uma matriz de proximidades, onde as linhas e colunas da matriz correspondem aos objetos submetidos à classificação. Se os objetos são caracterizados como padrões, ou pontos em um espaço $m$-dimensional, as proximidades ou similaridades podem ser representadas pelas distâncias entre os pares de pontos, e, a menos que uma medida significativa de distância entre os objetos tenha sido estabelecida, nenhuma análise de agrupamentos relevante pode ser realizada [13].

De acordo com [13], os problemas de classificação, que são tipos especiais de problemas de análise agrupamentos, podem ser categorizados em:

- Exclusivos e não exclusivos: uma classificação exclusiva é uma partição de um conjunto de padrões no qual cada objeto pertence a exatamente um subconjunto ou agrupamento, ao passo que uma classificação não exclusiva refere-se a uma situação na qual cada objeto pode estar associado a mais de um grupo;

- Intrínsecos e extrínsecos: são casos particulares dos problemas de classificação exclusivos. A classificação intrínseca utiliza somente a matriz de proximidades para realizar a categorização dos objetos. É nomeada como classificação não supervisionada por não empregar nenhum conhecimento anterior acerca dos elementos a agrupar. A classificação extrínseca, por outro lado, aplica além da matriz de proximidades uma coleção de objetos previamente classificados e o problema consiste em agrupar novos objetos, ainda não rotulados, com base nas informações obtidas por meio dos elementos já classificados. É qualificada como classificação supervisionada mais precisamente por utilizar conhecimento anterior acerca dos objetos submetidos ao agrupamento. Uma maneira de avaliar uma classificação não supervisionada consiste em verificar o quanto os rótulos dos grupos, associados aos objetos durante a categorização, relacionam-se aos rótulos dos objetos vinculados previamente. Por exemplo, considerando-se a existência de diversos índices de saúde pessoal, coletados para fumantes e não fumantes, uma classificação não supervisionada poderia dedicar-se a agrupar os indivíduos com base nos índices de saúde pessoal, além de procurar determinar se o hábito de fumar foi um fator preponderante ao desenvolvimento de diversas doenças. Neste mesmo contexto, uma classificação supervisionada poderia se ater em avaliar maneiras de determinar se o indivíduo é fumante ou não fumante, a partir dos índices de saúde pessoal;

- Hierárquicos e não-hierárquicos: por fim, os problemas de classificação exclusivos e intrínsecos são subdivididos em hierárquicos e não-hierárquicos, conforme o tipo de estrutura imposta sobre os dados. Enquanto que as classificações hierárquicas correspondem a uma sucessão de partições aninhadas, as categorizações não-hierárquicas representam partições disjuntas dos objetos.

Um agrupamento não-hierárquico consiste meramente na organização de uma coleção de padrões em um conjunto de grupos disjuntos, isto é, sem interseção, de modo que cada padrão esteja precisamente em um grupo. Fundamentalmente, dado um conjunto $D$, compreendendo $n$ padrões, e um número de grupos $K$, com $(K \leq n)$, um método de particionamento não-hierárquico procura organizar os padrões em $K$ partições, de sorte que os elementos que pertencem a um dado grupo sejam similares quando comparados entre si e dissimilares quando comparados a objetos que pertecem a um dado grupo distinto [12]. O valor $K$ pode ou não ser conhecido e um critério de avaliação, classificado como local ou global, deve ser adotado. Um critério global representa cada grupo por meio de um protótipo e associa os padrões aos grupos segundo o protótipo mais similar, enquanto que um critério local constitui os grupos utilizando a própria estrutura dos dados, a exemplo de grupos que podem ser estabelecidos a partir da identificação de regiões de alta densidade no espaço de padrões, ou por meio da associação de um padrão e seus $K$ vizinhos mais próximos a um mesmo grupo em particular [13].

## 4. Agrupamento de documentos

Um procedimento de agrupamento de objetos textuais tem como objetivo realizar o particionamento não-hierárquico de uma coleção de documentos em um determinado número de grupos, de modo que documentos similares são associados ao mesmo grupo enquanto que documentos distintos são distribuídos em diferentes grupos. Esta é uma operação que determina a estrutura subjacente a um conjunto de objetos de dados e que possibilita uma organização e uma navegação eficientes em grandes coleções de arquivos textuais. O problema do agrupamento de documentos pode ser formalmente definido como: dados $i$) um conjunto de documentos $D = \{d_i\}, i = 1,...,n$; $ii$) um número desejado de grupos $K$; e $iii$) uma função objetivo que avalia a qualidade do agrupamento, deseja-se determinar uma associação $\gamma : D \rightarrow 1,...,K$ que minimiza (ou, em alguns casos, maximiza) a função objetivo. A função objetivo é normalmente definida em função da similaridade ou distância entre os documentos, e, em geral, demanda-se também que a associação $\gamma$ seja sobrejetiva, a fim de garantir que nenhum dos $K$ grupos esteja vazio [14, 15].

Devido à alta dimensionalidade dos textos, o agrupamento de documentos é tido como uma das difíceis tarefas da área de mineração de dados, requerendo o uso de algoritmos eficientes que sejam capazes de manipular conjuntos constituídos por elementos de elevadas proporções. O processo padrão de agrupamento de documentos é usualmente constituído das seguintes etapas [16, 17, 18, 19, 20, 21, 22]:

- Pré-processamento: como os documentos que serão agrupados estão em um formato não-estruturado, etapas de pré-processamento devem ser realizadas antes que as técnicas de agrupamento possam ser efetivamente aplicadas. O pré-processamento inclui atividades de:
  - Identificação de termos: tem como objetivo selecionar os termos que serão utilizados para representar os documentos;
  - Lematização das palavras: envolve a eliminação das variações morfológicas de uma mesma palavra através da identificação do seu radical. Por exemplo, ”computador” e ”computação” são convertidas para forma base ”comput”. De modo similar, as palavras ”programação” e ”programar” seriam substituídas por ”program”;
  - Remoção de termos irrelevantes: tem como intuito eliminar palavras não relevantes para análise do texto, justamente por não representarem a sua ideia principal. Fazem parte da lista de termos não relevantes: preposições, pronomes, artigos, advérbios, além de outras classes de palavras auxiliares.

- Seleção de características e do modelo de representação dos documentos: consiste na representação do documento em um formato adequado à aplicação dos métodos de agrupamento. A forma de representação mais comum corresponde ao modelo espaço vetorial, no qual cada documento é tratado como um *bag-of-words* que utiliza as palavras como medida para identificar a similaridade entre os documentos. Recorrendo a este modelo, cada documento $d_i$ é considerado um ponto em espaço vetorial $m$-dimensional, $d_i = (w_{i1}, w_{i2}, \ldots, w_{im})$, $i = 1, \ldots, n$, no qual a dimensão $m$ corresponde ao número de termos distintos da coleção de documentos. Cada componente de $d_i$ representa um termo da coleção que pode ou não estar presente no documento, e o valor de cada componente depende do grau de relacionamento entre o termo e o documento que possivelmente o contém. Um dos esquemas mais utilizados para medir o relacionamento entre os termos e os documentos é o *tf-idf (Term Frequency-Inverse Document Frequency)*, calculado como $w_{ij} = n_{ij} \log_2 \left(\frac{n}{n_j}\right)$, onde $n_{ij}$ denota a frequência do termo, ou seja, quantas vezes o termo $t_j$ ocorre no documento $d_i$, $n_j$ corresponde ao número de documentos nos quais o termo $t_j$ aparece e $n$ representa o número de documentos da coleção;

- Seleção da medida de dissimilaridade: é um aspecto essencial do processo de agrupamento, pois quando formulado como um problema de otimização terá como função objetivo uma expressão que será dependente da medida de dissimilaridade. A dessemelhança entre dois documentos é determinada por meio de uma das diversas medidas de dissimilaridade baseadas nos vetores de características que os representam, a exemplo da distância Euclidiana, da distância do coseno, do coeficiente de Jaccard estendido e do coeficiente de correlação de Pearson;

- Aplicação do algoritmo de agrupamento: tem como resultado a geração dos agrupamentos baseados na medida de similaridade e no modelo de representação selecionados. O agrupamento originado pode ser rígido, que consiste em uma partição de dados entre os grupos, ou *fuzzy*, no qual cada padrão faz parte de cada um dos grupos, porém com diferentes graus de pertinência;

- Avaliação do agrupamento: consiste na aplicação de um critério de validação com o objetivo de avaliar a qualidade dos agrupamentos obtidos pelo método de agrupamento selecionado. Os critérios de validação podem ser classificados como externos ou internos. Os critérios externos, a exemplo da entropia, da pureza e do índice Rand, avaliam a performance comparando a estrutura do agrupamento resultante com algum conhecimento anterior, ao passo que os critérios internos, tais como o coeficiente silhueta, o índice Davies-Bouldin e o índice Dunn, permitem comparar diferentes conjuntos de grupos sem nenhuma referência a qualquer informação externa;

- Seleção de descritores para agrupamento: tem como objetivo selecionar os descritores que auxiliarão na interpretação dos resultados obtidos pelos métodos de agrupamento de documentos. Consiste em uma atividade importante, pois tendo em vista que o agrupamento é em geral utilizado em atividades exploratórias para descoberta de conhecimento, torna-se necessário indicar o significado de cada grupo de forma que os interessados (usuários ou aplicações) possam interagir com os agrupamentos de maneira mais intuitiva. Um dos métodos de identificação de descritores consiste em computar o coeficiente de correlação dos termos que pertencem ao grupo, ordená-los em ordem decrescente e em seguida selecionar os $n$ termos mais bem posicionados como os descritores. O coeficiente de correlação, que afere o relacionamento entre o termo $t$ e o grupo $c_k$ e que é apropriado em assinalar as palavras altamente indicativas da pertinência em relação à uma dada categoria, é representado por intermédio da expressão $C_t^{(k)} = \frac{VP \cdot VN - FN \cdot FP}{\sqrt{(VP+FN)\cdot(FP+VN)\cdot(VP+FP)\cdot(FN+VN)}}$, na qual $VP = verdadeiro\ positivo$ ou número de documentos que pertencem a $c_k$ e que contêm o termo $t$, $FP = falso\ positivo$ ou número de documentos que não pertencem a $c_k$ e que contêm o termo $t$, $FN = falso\ negativo$ ou número de documentos que não pertencem a $c_k$ e que não contêm o termo $t$ e $VN = verdadeiro\ negativo$ ou número de documentos que pertencem a $c_k$ e que não contêm o termo $t$.

## 5. Algoritmos de agrupamento

### 5.1 *K-means*

O algoritmo *K-means* e suas variações têm como intuito particionar um conjunto de $n$ objetos em $K$ grupos de modo que a similaridade entre os elementos que fazem parte de um mesmo grupo seja alta e que a semelhança entre os objetos que pertencem a grupos distintos seja baixa. Seja $X = \{x_i\}, i = 1, ..., n$ uma coleção de $n$ objetos com $m$-dimensões que devem ser agrupados em um conjunto de $K$ grupos, $C = \{c_k\}, k = 1, ..., K$ tal que cada grupo $c_k$ contenha $n_k$ padrões. O algoritmo *K-means* determina uma partição de modo que o erro quadrático entre a média empírica do grupo e os objetos que pertencem ao mesmo seja minorada. Se $m^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i^{(k)}$, com $x_i^{(k)}$ correspondendo ao $i$-ésimo padrão pertencente ao grupo $c_k$, representar a média do grupo $c_k$, então o erro quadrático entre $m^{(k)}$ e os objetos que pertencem ao grupo $c_k$ será expresso por $e_k^2 = \sum_{i=1}^{n_k} \parallel x_i^{(k)} - m^{(k)} \parallel^2$. Minimizar a função objetivo descrita pelo erro quadrático constitui um problema *NP*-difícil mesmo para $K = 2$. Desta forma, o procedimento *K-means*, que é classificado como um método guloso, é capaz de convergir somente para míminos locais, muito embora alguns estudos demonstrem que este algoritmo pode, com elevada probabilidade, convergir para ótimos globais sobretudo em situações nas quais os grupos de objetos apresentam-se bem separados. O método *K-means* inicia a sua execução com uma partição preliminar constituída de $K$ grupos e interativamente associa os padrões aos grupos de modo a reduzir o erro quadrático. Dado que o erro quadrático invariavelmente diminui em função do incremento do número de grupos $K$, o seu valor é verdadeiramente minimizado somente quando a quantidade de grupos permanece inalterada [23, 13, 11, 1, 24].

O método de agrupamento *K-means*, foi proposto originalmente por [25], e, a despeito de ter sido estabelecido há mais de 50 anos, consiste em um dos mais populares expedientes de particionamento, haja vista que a sua eficiência e os seus bons resultados experimentais contribuem para que este algoritmo seja assiduamente aplicado a problemas de classificação não supervisionada, a exemplo dos trabalhos de [26, 27, 28, 29, 30, 31], entre outros. De acordo com [1, 23], a estratégia de agrupamento *K-means* funciona conforme descrito a seguir. Inicialmente, $K$ objetos são aleatoriamente selecionados para representar as médias ou centróides dos grupos. Para cada um dos elementos restantes, não escolhidos como centróides iniciais, o algoritmo associa o objeto ao grupo mais próximo, baseado na medida de distância entre o objeto e a média do grupo, a qual corresponde ao vetor representado pela média dos valores de cada componente dos objetos designados ao grupo. Uma vez que todos os objetos tenham sido incorporados aos seus respectivos grupos, as médias dos $K$ grupos são recalculadas, com este processo sendo repetido até que uma condição de convergência seja satisfeita.

O procedimento *K-means* comporta-se particularmente bem quando os grupos de objetos são densos e bastante separados uns dos outros. O método é relativamente escalável e eficiente para o processamento de conjuntos de dados extensos, por apresentar complexidade de ordem linear, muito embora, e conforme já destacado, frequentemente resulte em mínimos locais. É uma estratégia que apresenta a desvantagem de necessitar que o número de grupos $K$ tenha que ser especificado com antecedência, não é adequado para identificar grupos de apresentação não convexa ou de dimensões excessivamente distintas, além de ser sensível a ruídos ou objetos discrepantes, que mesmo em número reduzido podem substancialmente influenciar na determinação dos centróides dos grupos [12].

Neste trabalho, o agrupamento de documentos com a aplicação do algoritmo *K-means* foi realizado observando as principais características do método original, contudo, algumas alterações, determinadas com base nos estudos de [23] e [32], foram estabelecidas:

- Cada documento $d_i$ foi representado como um ponto no espaço $m$-dimensional e cada componente de $d_i$ foi computada conforme a expressão $w_{ij} = n_{ij} \log_2 \left(\frac{n}{n_j}\right)$;
- A média ou centróide de cada grupo foi determinada pela expressão $m^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i^{(k)}$;
- A função objetivo a minimizar, que foi definida como o somatório da distância média dos documentos aos centróides dos grupos, foi calculada de acordo com a equação $f = \left[\sum_{k=1}^{K} \frac{1}{n_k}\sum_{i=1}^{n_k} d(m^{(k)}, d_i^{(k)})\right] \frac{1}{K}$;
- No intuito de procurar evitar que o algoritmo originasse soluções que correspondessem a mínimos locais, o método foi modificado para realizar dez inicializações com a posterior seleção do melhor resultado, de acordo com o critério definido pela função objetivo, ao final da execução;
- A fim de prevenir a ocorrência de grupos vazios, as soluções originadas ao final de cada execução do algoritmo foram submetidas a um procedimento que associava aos grupos sem documentos, objetos oriundos dos grupos com maior variação interna, a qual era determinada pela distância média dos elementos pertinentes a um grupo ao centróide do mesmo, segundo a expressão $v_k = \frac{1}{n_k}\sum_{i=1}^{n_k} d(m^{(k)}, d_i^{(k)})$, onde $n_k$ correspondia ao número de documentos associados ao $k$-ésimo grupo, $d$ consistia em uma função que estabelecia a distância entre os documentos, $m^{(k)}$ denotava o centróide ou média do $k$-ésimo grupo e $d_i^{(k)}$ representava o $i$-ésimo documento do $k$-ésimo grupo;
- O critério de interrupção do algoritmo foi modificado para que a execução fosse descontinuada quando o número máximo de dez iterações fosse alcançado ou quando não houvesse mais alterações nos centróides até então determinados.

**Algoritmo 1:** Algoritmo *K-means* aplicado ao agrupamento de documentos

```
Defina o número de inicializações NI;
Defina o número de máximo de iterações NMI;
para i ← 1 até NI faça
    ie ← 0;
    nmi ← 0;
    Selecione K documentos como centróides iniciais;
    enquanto ie = 0 faça
        Construa K grupos associando cada documento
          ao centróide mais próximo;
        Recalcule o centróide de cada grupo;
        nmi ← nmi + 1;
        se os centróides não se modificaram ou
          nmi = NMI então
            ie ← 1;
    Verifique se há grupos vazios na partição obtida
      pela inicialização de ordem i e os preencha com
      documentos originados dos grupos de maior
      variação interna;
    Calcule o valor da função objetivo da partição
      obtida pela inicialização de ordem i e a armazene
      em uma lista de soluções LS;
Selecione de LS a melhor partição e a defina como
  solução para o problema;
```

As operações executadas pelo método *K-means* observando seus aspectos característicos e as modificações propostas podem ser representadas pelo algoritmo 1.

## 5.2 Pesquisa harmônica

A pesquisa harmônica representa uma meta-heurística baseada em população que imita o processo de improvisão musical realizado por instrumentistas que simulam acordes em seus instrumentos no intuito de alcançar um estado harmônico perfeito [33]. Constitui um método que tem sido frequentemente abordado na literatura e que tem sido empregado na resolução de problemas de diversas naturezas, a exemplo de projetos de redes de abastecimento de água [34], projetos de expansão de redes de distribuição de energia elétrica [35], determinação dos parâmetros de corte utilizados na fabricação de artefatos de aço inoxidável [36], projetos de redes de distribuição de energia elétrica [37], operação de sistemas de energia hidroelétrica [38], ajuste de parâmetros de amortecedores de massa [39], entre outros. A pesquisa harmônica procura imitar o fenômeno natural representado pelo comportamento dos músicos que individualmente emitem acordes por meio de seus instrumentos e que cooperam entre si com o intuito de alcançar, sob o aspecto estético, uma excelente harmonia. Corresponde a um algoritmo capaz de explorar o espaço de busca de um determinado dado em um ambiente de otimização paralela e que possui muitas características que o tornam um método preferível não somente quando utilizado isoladamente, mas também quando associado a outras meta-heurísticas [33].

A analogia entre a improvisão musical e a resolução de problemas de otimização, conforme proposta pela pesquisa harmônica, pode ser descrita como se segue: *i*) cada músico corresponde a cada variável de decisão; *ii*) cada intervalo de acordes suportado pelo instrumento musical corresponde a cada intervalo de valores que a variável de decisão pode assumir; *iii*) o estado da harmonia musical em um determinado momento corresponde à representação vetorial da solução do problema em uma determinada iteração; *iv*) A percepção estética da harmonia representa a avalição da função objetivo. Se um dado problema de otimização for definido como minimizar $f(a)$ sujeito a $a_i \in A_i$, $i = 1,2,\ldots,n$, onde $f(a)$ representa uma função objetivo; $a$ corresponde ao conjunto que abrange cada variável de decisão $a_i$; $A_i$ expressa o conjunto que delimita o intervalo de valores de cada variável de decisão, com $LI_i \leq A_i \leq LS_i$ para $1 \leq i \leq n$, se $A_i$ é contínua, ou $A_i \in \{A_{i,1}, A_{i,2}, \ldots, A_{i,K_i}\}$, se $A_i$ é discreta; e $n$ denota a quantidade de variáveis de decisão, então os parâmetros da pesquisa harmônica que precisam ser definidos são: *i*) o tamanho da memória harmônica *HMS*, que corresponde ao número de soluções candidatas presentes na memória; *ii*) a proporção de consideração de soluções da memória harmônica *HMCR*, com $HMCR \in [0,1]$; *iii*) a proporção de ajuste de acorde *PAR*, com $PAR \in [0,1]$; *iv*) O critério de parada, em geral representado pelo número máximo de improvisações [33, 40].

A etapa de inicialização da memória harmônica *HM* consiste no preenchimento de uma matriz aumentada $HMS \times (n+1)$ que a representa, e que é interativamente atualizada durante o processo de otimização [40].

$$HM = \begin{bmatrix} a_1^1 & a_2^1 & \ldots & a_n^1 & f(a^1) \\ a_1^2 & a_2^2 & \ldots & a_n^2 & f(a^2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_1^{HMS} & a_2^{HMS} & \ldots & a_n^{HMS} & f(a^{HMS}) \end{bmatrix}$$

Os elementos armazenados em *HM* retratam as variáveis de decisão do problema e seus respectivos valores de função objetivo, inicialmente originados de forma aleatória segundo uma distribuição uniforme sobre o intervalo $[LI_i, LS_i]$, com $1 \leq i \leq n$, por meio da expressão $a_i^j = LI_i + r(LS_i - LI_i)$, $j = 1,2,\ldots,HMS$, onde $r$ corresponde a um número randômico uniformemente selecionado a partir do intervalo $[0,1]$. A improvisação de uma nova harmonia representa a principal operação executada pelo algoritmo e constitui o elemento essencial sobre o qual a pesquisa harmônica encontra-se fundamentada. Nesta etapa o método origina um novo vetor harmônico $a' = (a'_1, a'_2, a'_3, \ldots, a'_n)$, a partir de soluções oriundas da memória harmônica, com base em soluções determinadas aleatoriamente ou ainda segundo soluções modificadas por meio do operador de ajuste de acorde. Ao considerar soluções provindas da memória, os valores do novo vetor harmônico são aleatoriamente obtidos dos valores armazenados em *HM*, com probabilidade $HMCR \in [0,1]$. Desta

forma, o valor da variável de decisão $a'_1$ é selecionado de $(a^1_1, a^2_1, a^3_1, \ldots, a^{HMS}_1)$, o valor da variável $a'_2$ é originado de $(a^1_2, a^2_2, a^3_2, \ldots, a^{HMS}_2)$, e assim sucessivamente para todas as demais variáveis de decisão $(a'_3, a'_4, a'_5, \ldots)$. Quando as novas soluções não são oriundas da memória harmônica, os valores das variáveis de decisão que as constituem são aleatoriamente selecionados de acordo com o intervalo admissível, $a'_i \in A_i$. Esta situação, que ocorre com probabilidade $(1 - HMCR)$ e é referida como consideração aleatória, promove uma maior diversificação dos resultados por orientar o método a explorar um espaço de soluções mais abrangente, de modo que soluções classificadas como ótimas globais possam ser alcançadas [40, 33].

De acordo com [33], as etapas do método que derivam soluções da memória harmônica ou que as originam de maneira aleatória podem ser representadas por meio da expressão $a'_i \leftarrow a'_i \in \{a^1_i, a^2_i, a^3_i, \ldots, a^{HMS}_i\}$ com probabilidade $HMCR$, ou por intermédio da expressão $a'_i \leftarrow a'_i \in A_i$ com probabilidade $(1 - HMCR)$. No intuito de proporcionar uma busca adicional por bons resultados sobre o espaço de pesquisa das soluções, as variáveis de decisão $a'_i$ que compõem o vetor harmônico $a' = (a'_1, a'_2, a'_3, \ldots, a'_n)$, podem ser individualmente examinadas e submetidas, com probabilidade $PAR \in [0,1]$, a uma operação denominada ajuste de acorde, a qual modifica o conteúdo da variável de decisão $a'_i$ por meio da equação $a'_i = a'_i \pm rbw$, onde $r$ corresponde a um número randôminco $\in [0,1]$ e $bw$ representa um intervalo arbritário empregado com o objetivo de melhorar a performance do algoritmo. O valor de $bw$, que pode ser discreto ou contínuo conforme o problema de otimização, determina a extensão das alterações ou deslocamentos que podem ocorrer sobre os valores das variáveis de decisão que constituem o novo vetor. Com o propósito de atualizar a memória harmônica com o vetor harmônico modificado $a' = (a'_1, a'_2, a'_3, \ldots, a'_n)$, o valor da função objetivo $f(a')$ é calculado e comparado com o pior vetor harmônico existente na memória. Se, com relação à função objetivo, o novo vetor for melhor do que a pior solução armazenada, então esta será substituída pela solução modificada. Caso contrário, o novo vetor harmônico é ignorado. O processo iterativo do método é interrompido quando o número máximo de improvisações é alcançado, e, por fim, o melhor vetor existente na memória é selecionado e considerado a melhor solução para o problema sob investigação.

De acordo com [33], as operações executadas pelo método de pesquisa harmônica podem ser representadas por meio do algoritmo 2.

**Algoritmo 2:** Algoritmo pesquisa harmônica

```
Defina uma função de avaliação f(a), a = (a_1, a_2, ..., a_n);
Defina HMCR, PAR e HMS;
Defina o número máximo de improvisações NI;
Inicialize a memória harmônica HM;
Defina o limite inferior das variáveis de decisão LI;
Defina o limite superior das variáveis de decisão LS;
enquanto número de improvisações ≤ NI faça
    enquanto a'_i ≤ número de variáveis faça
        se r ∈ [0,1] ≤ HMCR então
            Selecione para a'_i um valor proveniente de HM;
            se r ∈ [0,1] ≤ PAR então
                Ajuste o valor de a'_i por:
                a'_i ← a'_i ± r ∈ [0,1]bw;
        senão
            Selecione para a'_i um valor aleatório:
            a'_i ← LI + r ∈ [0,1](LS − LI);
    se f(a') ≤ f(pior solução de HM) então
        Substitua a pior solução de HM por a';
Selecione de HM o melhor vetor harmônico e o defina como solução para o problema;
```

Um trabalho realizado por [32] apresentou um novo algoritmo de agrupamento de documentos baseado na metaheurística pesquisa harmônica integrada ao método *k-means*, que tinha como intuito obter agrupamentos de melhor qualidade a partir do emprego do poder exploratório da pesquisa harmônica unido à capacidade de refinamento do *k-means*.

No trabalho conduzido por [32], o problema do agrupamento de documentos foi abordado como uma questão de otimização que teve como objetivo identificar os melhores centróides dos grupos ao invés de determinar a melhor partição dos elementos. Para este fim definiu-se a qualidade dos grupos como função objetivo e utilizou-se o algoritmo de pesquisa harmônica no intuito de otimizar esta função. Os objetos foram representados aplicando-se o modelo espaço vetorial multidimensional no qual cada documento $d_i = (w_{i1}, w_{i2}, \ldots, w_{im})$ era considerado um vetor no espaço de termos constituído por $m$ termos distintos, cujo valor de cada característica era determinado por meio da expressão *tf-idf*, e onde cada possível solução de agrupamentos foi definida como um vetor de centróides de tamanho $n$, com $n$ representando o número de documentos. Cada elemento do vetor solução era retratado por um inteiro no intervalo $[1, K]$, com $K$ denotando o número de grupos, que indicava a qual grupo o documento pertencia. O espaço de pesquisa foi constituído por todas as permutações de tamanho $n$ do conjunto $\{1, ..., K\}$ satisfazendo às restrições de que o algoritmo deveria alocar cada documento a exatamente um grupo e de que nenhum grupo poderia resultar vazio. Na modelagem proposta, cada linha da memória harmônica foi constituída por um vetor de inteiros de $n$ posições no qual a $i$-ésima posição continha o número do grupo associado ao $i$-ésimo documento, conforme demonstrado, por exemplo, pelo elemento $a = (3,2,1,1,3,3,2,2)$, que poderia representar uma possível solução para o problema do agrupamento de oito documentos em três grupos.

Na etapa de inicialização do algoritmo, a memória harmônica era preenchida com um conjunto de vetores de soluções, determinados de modo aleatório, de forma que cada linha da

memória correspondia a um agrupamento específico de documentos em que o $i$-ésimo elemento era selecionado a partir de uma distribuição uniforme sobre o conjunto $\{1,\ldots,K\}$. Na etapa de improvisação o número do grupo associado a cada documento no novo vetor solução $a'$ era originado da memória harmônica com probabilidade $HMCR$ e com probabilidade $1-HMCR$ aleatoriamente selecionado do conjunto $\{1,\ldots,K\}$. Após a determinação da nova solução, o ajuste de acorde era aplicado com probabilidade $PAR$, calculada conforme a expressão $PAR = PAR_{min} + \frac{PAR_{max}-PAR_{min}}{NI} ni$, onde $PAR_{min}$ e $PAR_{max}$ representavam nesta ordem os valores mínimo e máximo da proporção de ajuste de acorde, $NI$ consistia no número total de improvisações e $ni$ representava o número da improvisação em execução. A probabilidade $PAR$ determinava a razão com que um grupo distinto daquele originado da memória harmônica era relacionado a um documento. Tendo em vista que a pesquisa harmônica e suas variações haviam sido inicialmente desenvolvidas para problemas de otimização que envolviam variáveis de decisão contínuas, e que o novo algoritmo utilizava uma representação de soluções constituída por variáveis discretas, os autores propuseram a utilização de dois parâmetros de ajuste de acorde: $PAR_1 = 0,6PAR$ e $PAR_2 = 0,3PAR$. Para cada documento $d_i$, o grupo selecionado da memória harmônica era, com probabilidade $PAR_1$, substituído pelo grupo cujo centróide estivesse mais próximo de $d_i$, conforme determinado pela expressão $a'(i) = min\ d(d_i, m^{(k)}), k \in \{1,\ldots,K\}$, onde $d$ consiste em uma função que estabelecia a distância entre os documentos e $m^{(k)}$ correspondia ao elemento que representa o centroíde ou média do $k$-ésimo grupo. Com probabilidade $PAR_2$ o grupo associado ao documento $d_i$ era substituído por um novo grupo selecionado aleatoriamente com base na distribuição $p_k = \frac{d_{max}-d(d_i,m^{(k)})}{NF}\left(1-\frac{ni}{NI}\right), k \in \{1,\ldots,K\}$, onde $p_k$ correspondia à probabilidade do grupo $k$ ser selecionado como o novo grupo, $NF = Kd_{max} - \sum_{k=1}^{K} d(d_i^{(k)}, m^{(k)})$, $d_{max} = max_k\ d(a', m^{(k)})$, $NI$ consistia no número total de improvisações e $ni$ representava o número da improvisação em execução. Na etapa de avaliação das soluções, os $K$ centróides associados às partições representadas pelas linhas da memória harmônica eram determinados pela média dos documentos que pertenciam a cada agrupamento, por intermédio da expressão $m^{(k)} = \frac{1}{n_k}\sum_{i=1}^{n_k} d_i^{(k)}$, onde $n_k$ denotava o número de documentos que pertenciam ao grupo $K$ e $d_i^{(k)}$ correspondia ao $i$-ésimo documento presente no grupo $K$. A função objetivo, que consistia na distância média dos documentos ao centróide do grupo ao qual pertenciam e que tinha como propósito maximizar a similaridade intra-grupos (minimizando a distância intra-grupos) e minimizar a similaridade inter-grupos (maximizando a distância entre os grupos), era computada para cada linha da memória harmônica por meio da equação $f = \left[\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k} d(m^{(k)}, d_i^{(k)})\right]\frac{1}{K}$ onde $K$ significava o número de grupos, $n_k$ correspondia ao número de documentos associados ao $k$-ésimo grupo, $d$ consistia em uma função que estabelecia a distância entre os documentos, $m^{(k)}$ denotava o centróide do $k$-ésimo grupo e $d_i^{(k)}$ representava o $i$-ésimo documento do $k$-ésimo grupo. Ao final da etapa de avaliação, o valor da função objetivo da nova solução originada na etapa de improvisação era comparado com os valores de aptidão das soluções armazenadas na memória, com a substituição da solução menos capaz pela solução recém-calculada nos casos em que esta apresentava melhor qualidade. O critério de parada do método correspondia ao alcance do número máximo de improvisações ou à não modificação da aptidão média das soluções por um valor superior a um dado $\alpha$ dentro de um número de iterações estimado previamente [32].

O estudo realizado por [32] propôs ainda a hibridização do novo método com o algoritmo *k-means* a fim de aproveitar a sua capacidade de busca nas proximidades de uma solução, no sentido de refinar o resultado obtido pela pesquisa harmônica com a consequente redução do tempo de convergência do processo.

Neste trabalho, o agrupamento de documentos com o emprego da meta-heurísitica pesquisa harmônica foi realizado levando em consideração muitas das particularidades apresentadas em [32], entretanto, algumas alterações, estabelecidas com base nos estudos de [41] e [42], foram propostas:

- O centróide de cada grupo foi representando por um dos objetos da coleção e não pela média dos objetos que pertenciam ao grupo. Este modelo permitia a utilização de uma matriz de proximidades $n \times n$, que era empregada para armazenar as dissimilaridades entre os $n$ documentos no espaço $m$-dimensional, e que previnia o recálculo dos centróides originalmente exigido pelo reposicionamento dos objetos entre os grupos, durante as improvisações realizadas pelo algoritmo;

- Cada linha da memória harmônica correspondia a um vetor $K$-dimensional, cujas posições podiam ser ocupadas por qualquer um dos $n$ objetos que representavam os centróides dos grupos;

- A etapa de inicialização da memória harmônica foi modificada para que $K$ objetos fossem aleatoriamente selecionados como centróides, com a distribuição posterior dos demais elementos entre os grupos a partir da menor distância entre o objeto e um dado centróide;

- A operação de ajuste de acorde, aplicada sobre o novo vetor solução $a'$ originado da memória harmônica, foi alterada para sugerir um novo centróide de grupo na $i$-ésima posição de $a'$ conforme as expressões: $i) a'_i \leftarrow a'_i + w \in \{1,\ldots,n-1\}$ se $a'_i = 1$; $ii) a'_i \leftarrow a'_i - w \in \{1,\ldots,n-1\}$ se $a'_i = n$; $iii) a'_i \leftarrow a'_i + w \in \{1,\ldots,n - a'_i\}$ se $1 < a'_i < n$ e o ajuste fosse positivo; $iv) a'_i \leftarrow a'_i - w \in \{1,\ldots,a'_i - 1\}$ se $1 < a'_i < n$ e o ajuste fosse negativo;

- A etapa de hibridização com o método *K-means* foi suprimida;

- O critério de suspensão do algoritmo foi modificado para que a interrupção acontecesse quando o número máximo de improvisações fosse alcançado ou quando o número máximo de improvisações consecutivas sem melhoria na função objetivo fosse atingido.

Cumpre salientar que as modificações sugeridas na operação de ajuste de acorde eventualmente originavam soluções com centróides em duplicidade, as quais eram desconsideradas por não acatarem formalmente à restrição de que um objeto não poderia pertencer simultâneamente a mais de um grupo. As operações executadas pelo método de pesquisa harmônica, observando os aspectos presentes em [32] e as modificações propostas podem ser representadas pelo algoritmo 3.

**Algoritmo 3:** Algoritmo pesquisa harmônica aplicado ao agrupamento de documentos

```
1  Defina f(a);
2  Defina HMCR, PAR_max, PAR_min e HMS;
3  Defina o número de máximo de improvisações NI;
4  Defina o número de máximo de improvisações sem
     melhoria NISM;
5  ni ← 1;
6  nism ← 1;
7  Inicialize a memória harmônica HM;
8  enquanto ni ≤ NI e nism ≤ NISM faça
9      enquanto a'_i ≤ número de variáveis faça
10         se r ∈ [0,1] ≤ HMCR então
11             Selecione para a'_i um valor proveniente de
                 HM;
12             PAR ← PAR_min + ((PAR_max − PAR_min)/NI) ni;
13             se r ∈ [0,1] ≤ PAR então
14                 se a'_i = 1 então
15                     a'_i ← a'_i + w ∈ {1,...,n−1};
16                 se a'_i = n então
17                     a'_i ← a'_i − w ∈ {1,...,n−1};
18                 se 1 < a'_i < n então
19                     se r ∈ [0,1] > 0,5 então
20                         a'_i ← a'_i + w ∈ {1,...,n − a'_i};
21                     senão
22                         a'_i ← a'_i − w ∈ {1,...,a'_i − 1};
23         senão
24             a'_i ← w ∈ {1,...,n};
25     se a' não foi rejeitado então
26         se f(a') ≤ f(pior solução de HM) então
27             Substitua a pior solução de HM por a';
28     se f(a') > f(pior solução de HM) e a' não foi
         rejeitado então
29         nism ← nism + 1;
30     senão
31         nism ← 0;
32     ni ← ni + 1;
33 Selecione de HM o melhor vetor harmônico e o defina
     como solução para o problema;
```

## 5.3 Algoritmo genético

Os algoritmos genéticos são uma ramificação dos algoritmos evolucionários e podem ser definidos como técnicas de busca fundamentadas numa metáfora do processo biológico de evolução natural. Nos algoritmos genéticos, populações de indivíduos são concebidas e submetidas aos operadores genéticos, que utilizam uma função de avaliação para caracterizar a qualidade de cada indivíduo como solução para o problema abordado, gerando um processo de evolução natural que eventualmente deverá resultar em um indivíduo que representará uma boa solução [43].

A codificação da informação em cromossomos se constitui em um aspecto crucial do algoritmo genético, visto que representa juntamente com a função de avaliação o elemento que associa o método ao problema a ser resolvido. Se a codificação for realizada de maneira adequada, então esta já incluirá as particularidades do problema, evitando a execução de testes de viabilidade para cada uma das soluções originadas. Os cromossomos presentes na população do algoritmo genético, que podem ser interpretados como pontos no espaço de pesquisa das soluções candidatas, são representados por meio de sequências de bits, nas quais cada posição apresenta dois possíveis valores: 0 e 1. O algoritmo genético analisa as populações de cromossomos, sucessivamente substituindo uma população por outra com base na função de avaliação [43, 44].

De acordo com [45, 44, 43], os operadores genéticos presentes mesmo nas representações mais elementares dos algoritmos genéticos podem ser descritos conforme a seguir:

- Seleção: é o operador responsável pela designação dos cromossomos que serão submetidos ao processo de reprodução. Por este mecanismo, quanto melhor for o cromossomo maior será a probabilidade de que o mesmo seja destacado para reproduzir;

- Recombinação ou *crossover*: este operador seleciona aleatoriamente uma posição de um par de cromossomos e permuta as subsequências situadas antes e após a posição escolhida, no intuito de originar dois descendentes que os substituirão. Por exemplo, as sequências 10000100 e 11111111 poderiam ser recombinadas após a terceira posição, originando como descendentes os cromossomos 10011111 e 11100100;

- Mutação: corresponde ao operador que aleatoriamente inverte alguns dos valores presentes na representação dos cromossomos, a exemplo da sequência 00000100 que poderia ser modificada na segunda posição a fim de resultar em 01000100. É um operador que pode

**Algoritmo 4:** Algoritmo genético simples

```
Estabeleça aleatoriamente uma população de soluções candidatas, denominada população original;
enquanto Não alcançar o critério de parada faça
    Estabeleça uma população nova, vazia;
    enquanto a população nova não estiver completamente preenchida faça
        Selecione aleatoriamente um par de cromossomos da população original de modo que indivíduos com maior aptidão tenham maior probabilidade de serem escolhidos;
        Recombine os cromossomos a fim de originar dois novos descendentes;
    Com probabilidade aleatória, realize sobre cada indivíduo da poulação nova uma operação de mutação;
    Substitua a população original pela população nova;
Selecione da população final o cromossomo com maior aptidão e o defina como solução para o problema;
```

ocorrer para cada uma das posições dos cromossomos que correspondem às soluções, mas em geral com uma probabilidade muito reduzida, da ordem de 0,005 a 0,01.

Segundo [46], o funcionamento de um algoritmo genético simples pode ser descrito conforme a seguir. Cada execução do algoritmo genético é chamada de geração. Usualmente, entre 50 e 500 gerações são suficientes, e, ao final de cada execução, frequentemente existirão um ou mais cromossomos com elevada relevância em relação à função de avaliação. Tendo em vista que a aleatoriedade exerce significativa influência sobre o algoritmo, duas execuções com diferentes inicializações de soluções candidatas geralmente irão resultar em comportamentos distintos [44].

O método evolução diferencial é apresentado como uma versão melhorada dos algoritmos genéticos que tem como propósito obter, com maior celeridade, soluções para problemas de otimização. Consiste em um método no qual a mutação corresponde a uma operação que origina um novo indivíduo por meio da adição da diferença ponderada entre duas soluções a uma terceira solução. A solução modificada, denominada solução doadora, é então associada com outra solução pré-determinada, a solução alvo, no intuito de originar a solução teste. Esta operação, que tem como objetivo aumentar a diversidade das soluções mutadas por incorporar boas soluções de gerações anteriores, corresponde à recombinação. Caso a solução teste resulte em um valor de aptidão melhor do que aquele associado à respectiva solução alvo, a última será substituída pela primeira na próxima geração. Esta operação representa a seleção [47].

Um algoritmo genético hibridizado com a evolução diferencial discreta foi apresentado por [42]. Naquele trabalho os autores propuseram o uso das características do algoritmo genético em conjunto com as características da evolução diferencial discreta com o objetivo de melhorar o tempo de convergência do algoritmo genético na resolução do problema de agrupamento de documentos. No estudo, os documentos $D = \{d_i\}, i = 1,...,n$ foram representados como pontos no espaço vetorial $m$-dimensional, no qual $m$ correspondia ao número de termos distintos da coleção de textos e o valor de cada componente de $d_i$ expressava a frequência do termo no documento. Após retratar os documentos de forma vetorial, os autores calcularam as distâncias entre os textos e estabeleceram a construção de dois conjuntos numéricos retangulares: uma matriz de padrões $m \times n$, utilizada para representar a frequência dos $m$ termos nos $n$ documentos, e uma matriz de proximidades $n \times n$, empregada para armazenar as distâncias entre os $n$ documentos no espaço $m$-dimensional.

No estudo conduzido por [42], os cromossomos que peterciam à população de soluções foram representados por vetores de $K$ posições, cada uma das quais ocupadas por um inteiro no intervalo $[1,n]$, com $n$ determinando o número de documentos. O processo de inicialização da população consistia na seleção aleatória de $K$ documentos como centróides iniciais e na atribuição dos demais documentos aos grupos com base na menor distância entre o documento e um dado centróide. Neste modelo, diante, por exemplo, de um problema representado pelo agrupamento de 15 documentos em 5 grupos, um cromossomo da população de soluções candidatas poderia ser um vetor da forma $c = (2,5,8,10,1)$, que apresentaria 5 dimensões (número de grupos) ocupadas por valores oriundos do intervalo compreendido entre 1 e 15 (número de documentos). No trabalho de [42], o problema do agrupamento de documentos foi abordado como uma atividade de otimização, e a função objetivo ou de avaliação a minimizar, calculada para cada cromossomo da população, que foi definida como sendo a variância média dos documentos de um grupo ao centróide correspondente, era determinada por $f = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \left(d_i^{(k)} - m^{(k)}\right)^2$, onde $m^{(k)}$ correspondia ao documento selecionado como centro do grupo $K$, $d_i^{(k)}$ representava o $i$-ésimo documento pertencente ao grupo $K$ e $n_k$ denotava o número de documentos presentes no grupo $K$.

No algoritmo proposto por [42], no qual cada cromossomo correspondia a $K$ grupos distintos, havia a possibilidade de existirem centróides em comum no novo vetor solução resultante da recombinação. A fim de suplantar este impedimento, os autores propuseram a identificação e remoção dos centróides idênticos, de sorte que a operação de recombinação fosse realizada somente sobre os elementos não coincidentes de cada cromossomo. Por este método, se, por exemplo, dois cromossomos com 5 genes fossem representados pelos vetores $c = (1,4,6,7,9)$ e $c' = (5,11,10,8,1)$, então o elemento 1, comum aos vetores $c$ e $c'$, deveria ser suprimido antes que a recombinação fosse aplicada sobre os vetores modificados $c = (4,6,7,9)$ e $c' = (5,11,10,8)$. Posteriormente, aos vetores resultantes, eventualmente adicionavam-se os elementos

desmembrados dos vetores originais. De modo similar à recombinação, a mutação também poderia originar centróides em comum, desta forma os autores estabeleceram que esta operação deveria ser realizada levando em consideração somente valores ainda não presentes no cromossomo inicial. Por esta especificação, se, por exemplo, uma mutação tivesse que ser aplicada sobre o segundo gene do cromossomo $c = (1,4,6,7,9)$, o número 4 deveria ser substituído por um valor originado do conjunto $X = \{1,2,\ldots,n\} - \{1,4,6,7,9\}$.

O método híbrido descrito no trabalho de [42] resumia-se na inicialização da população e na execução sucessiva do algoritmo genético e da evolução diferencial discreta sobre as soluções representadas pelos cromossomos, por um dado número de iterações, de sorte que nas iterações de ordem ímpar o algoritmo genético era empregado e nas iterações de ordem par a evolução diferencial discreta era utilizada. No algoritmo genético, as operações de recombinação e mutação eram aplicadas sobre cromossomos escolhidos aleatoriamente da população original, com a consequente atualização da nova população com as soluções resultantes mais aptas em termos da função objetivo. Na evolução diferencial discreta, o melhor cromossomo da população vigente era submetido ao procedimento de mutação, seguido da operação de recombinação com um dado cromossomo da população original e da posterior substituição das soluções de qualidade inferior pelos cromossomos com melhores aptidões.

Neste trabalho, o agrupamento de documentos com o emprego do algoritmo genético híbrido com a evolução diferencial discreta é realizado levando em consideração muitas das particularidades apresentadas em [42], entretanto, algumas alterações, estabelecidas com base no estudo de [32], foram propostas:

- O valor de cada componente de $d_i$ foi determinado conforme a equação $w_{ij} = n_{ij}\log_2\left(\frac{n}{n_j}\right)$;
- A função objetivo a minimizar foi computada por meio da expressão $f = \left[\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k} d(m^{(k)}, d_i^{(k)})\right]\frac{1}{K}$;
- O critério de interrupção do método foi alterado para que a execução fosse descontinuada quando o número máximo de iterações sucessivas sem melhoria na função objetivo fosse alcançado.

As operações executadas pela rotina principal do método, observando os aspectos presentes em [42] e as modificações propostas podem ser representadas pelo algoritmo 5.

**Algoritmo 5:** Algoritmo genético híbrido com a evolução diferencial discreta aplicado ao agrupamento de documentos

```
Defina o tamanho da população de cromossomos TP;
Defina o número máximo de iterações sem melhoria NISM;
Defina a probabilidade de recombinação do algoritmo genético PRAG;
Defina a probabilidade de mutação do algoritmo genético PMAG;
Defina a probabilidade de recombinação da evolução diferencial PRED;
Defina a probabilidade de mutação da evolução diferencial PMED;
para i ← 1 até TP faça
    Selecione aleatoriamente K documentos dentre os n documentos e os considere como os K centros dos K grupos, de forma que cada coleção de centros represente um cromossomo da população inicial;
    para cada um dos n − K documentos faça
        Determine a distância do documento aos K centros;
        Associe o documento ao grupo cujo centro esteja mais próximo;
    Determine o valor de aptidão do cromossomo por meio da função de avaliação;
Considere a população aleatória inicial como a população original;
ie ← 0;
nism ← 0;
enquanto ie = 0 faça
    se i é ímpar então
        Execute o algoritmo genético;
    senão
        Execute a evolução diferencial discreta;
    se o cromossomo mais apto da nova população for melhor do que o cromossomo mais apto da população até então existente então
        nism ← 0;
    senão
        nism ← nism + 1;
    se nism ≥ NISM então
        ie ← 1;
Selecione da população final o cromossomo com maior aptidão e o defina como solução para o problema;
```

## 5.4 Método de agrupamento baseado no coeficiente de cobertura para bancos de dados textuais

O algoritmo C$^3$M *(Cover Coefficient-based Clustering Methodology)* consiste em um método que origina uma partição com base na designação de um conjunto de documentos caracterizados como centróides e na posterior associação dos demais documentos aos grupos inicialmente estabelecidos pelos centróides selecionados. O coeficiente de cobertura, conceito elementar sobre o qual método C$^3$M encontra-se fundamentado, propõe-se a [7]: *i*) identificar o relacionamento entre os documentos de uma coleção por meio do emprego de uma matriz nomeada como matriz *C*; *ii*) determinar o número de grupos presente na coleção de documentos; *iii*) selecionar

os centróides iniciais com a utilização do conceito de aptidão para centróide de grupo; *iv*) estabelecer os grupos com relação a C$^3$M, servindo-se dos conceitos *i*), *ii*) e *iii*) [7].

Se $D$ for matriz que representa uma coleção de documentos $\{d_i\}, i = 1,\ldots,m$, descritos pelos termos $\{t_i\}, i = 1,\ldots,n$, então a matriz $C$ consiste em uma matriz documento $\times$ documento cujas entradas $c_{ij}(1 \leq i,j \leq m)$ indicam a probabilidade de qualquer termo do documento $d_i$ ser selecionado a partir de um termo do documento $d_j$. A fim de que o conceito do coeficiente de cobertura possa ser utilizado, as entradas da matriz $D$, $d_{ij}(1 \leq i \leq m, 1 \leq j \leq n)$, devem satisfazer às seguintes condições: *i*) cada documento deve possuir pelo menos um termo; *ii*) cada termo deve estar presente em pelo menos um documento. Para determinação das entradas $c_{ij}$ da matriz $C$, é necessário a princípio selecionar aleatoriamente um termo $t_k$ do documento $d_i$ e usar este termo para tentar selecionar $d_j$, ou seja, verificar se $d_j$ encerra $t_k$. Em outras palavras, tem-se um ensaio probabilístico constituído por duas etapas e cada linha da matriz $C$ retrata os resultados deste experimento de dois estágios [7].

Se $s_{ik}$ indicar o evento que corresponde à seleção de $t_k$ a partir de $d_i$ no primeiro estágio e $s'_{jk}$ indicar o evento correspondente à seleção de $d_j$ a partir de $t_k$ no segundo estágio, então a probabilidade de ocorrência de $s_{ik}$ e $s'_{jk}$ pode ser representada pela expressão $P(s_{ik})P(s'_{jk})$, com $P(s_{ik}) = d_{ik}(\sum_{h=1}^{n} d_{ih})^{-1}$ e $P(s'_{jk}) = d_{jk}(\sum_{h=1}^{m} d_{hk})^{-1}$, para $1 \leq i,j \leq m$ e $1 \leq k \leq n$ [7].

No intuito de ilustrar este conceito, suponha a determinação de $c_{12}$ com base na matriz $D$ que se segue, a qual representa uma coleção constituída por cinco documentos (linhas) e seis termos distintos (colunas), cujas entradas preenchidas com 0 e 1 indicam, nesta ordem, a ausência ou a presença de um dado termo no documento. Segundo o modelo de probabilidade de dois estágios, para calcular $c_{12}$ seleciona-se um dos termos presentes em $d_1$ e então tenta-se selecionar $d_2$ a partir do resultado (termo) inicialmente obtido. No primeiro estágio, se o termo escolhido for $t_1$ ou $t_5$, então a probabilidade de selecionar $d_2$ será $\frac{1}{2}$, tendo em vista que $t_1$ e $t_5$ estão presentes somente em $d_1$ e $d_2$. De outra forma, se o termo $t_2$ for selecionado no primeiro estágio, então a probabilidade de selecionar $d_2$ no segundo estágio será $\frac{1}{4}$, considerando-se que $t_2$ ocorre em $d_1, d_2, d_4$ e $d_5$. No primeiro estágio, a probabilidade de se escolher um dos elementos de $\{t_1, t_2, t_5\}$ a partir de $d_1$ será $\frac{1}{3}$, enquanto que a esperança de escolher-se qualquer um dos outros termos será 0, uma vez que os elementos de $\{t_3, t_4, t_6\}$ não ocorrem em $d_1$. Por consequência, o valor de $c_{12}$ será determinado como: $c_{12} = \sum_{k=1}^{6} s_{1k}s'_{2k} = \frac{1}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{4} + 0\cdot 0 + 0\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{2} + 0\cdot 0 = 0{,}417$ [7].

$$D = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}$$

De acordo com [7], a matriz $C$ pode ser constituída a partir das matrizes denominadas $S$ e $S'$, de modo que $C = S \times S'^T$, e onde os elementos de $S$ e $S'$ são, respectivamente, representados por $s_{ik}$ e $s'_{jk}$, previamente descritos. Desta forma, utilizando as definições das matrizes $S$ e $S'$ os itens de $C$ podem ser originados por meio da expressão $c_{ij} = \sum_{k=1}^{n} s_{ik}s'^T_{kj}$, onde $s'^T_{kj} = s'_{jk}$. Esta expressão pode ainda ser reescrita como $c_{ij} = \alpha_i \sum_{k=1}^{n} d_{ik}\beta_k d_{jk}, 1 \leq i,j \leq m$, onde $\alpha_i$ e $\beta_k$ correspondem, respectivamente, ao inverso do somatório da $i$-ésima linha e da $k$-ésima coluna de $D$, conforme demonstrado pelas expressões $\alpha_i = \left[\sum_{j=1}^{n} d_{ij}\right]^{-1}, 1 \leq i \leq m$ e $\beta_k = \left[\sum_{j=1}^{m} d_{jk}\right]^{-1}, 1 \leq k \leq n$. Por este modelo e com base na matriz $D$, o valor de $c_{12}$ seria determinado por $c_{12} = \alpha_1 \cdot (d_{11}\cdot\beta_1\cdot d_{21} + d_{12}\cdot\beta_2\cdot d_{22} + d_{13}\cdot\beta_3\cdot d_{23} + d_{14}\cdot\beta_4\cdot d_{24} + d_{15}\cdot\beta_5\cdot d_{25} + d_{16}\cdot\beta_6\cdot d_{26}) = \frac{5}{12} = 0{,}417$, onde $\alpha_1 = \frac{1}{3}, \beta_1 = \frac{1}{2}, \beta_2 = \frac{1}{4}, \beta_3 = \frac{1}{2}, \beta_4 = \frac{1}{2}, \beta_5 = \frac{1}{2}$ e $\beta_6 = \frac{1}{3}$.

As seguintes propriedades são válidas para a matriz $C$: *i*) para $i \neq j, 0 \leq c_{ij} \leq c_{ii}$ e $c_{ii} > 0$; *ii*) $c_{i1} + c_{i2} + \cdots + c_{im} = 1$, ou seja, a soma dos elementos da $i$-ésima linha é igual a 1 para $1 \leq i \leq m$; *iii*) se nenhum dos termos de $d_i$ ocorre em qualquer outro documento, então $c_{ii} = 1$, caso contrário $c_{ii} < 1$; *iv*) se $c_{ij} = 0$, então $c_{ji} = 0$, e identicamente, se $c_{ij} > 0$, então $c_{ji} > 0$, mas em geral $c_{ij} \neq c_{ji}$; *v*) $c_{ii} = c_{jj} = c_{ij} = c_{ji}$ se e somente se $d_i$ e $d_j$ são idênticos [7].

As propriedades descritas indicam que se um documento possui muitos termos em comum com os demais documentos da coleção, então o documento apresentará muitos valores diferentes de zero nas entradas posicionadas fora da diagonal principal da linha correspondente. Neste caso, desde que a soma dos elementos de uma dada linha da matriz $C$ é igual a 1, o item equivalente à diagonal terá valor menor do que 1. De modo contrário, ou seja, se o documento possui poucos termos em comum com o restante da coleção, então haverá muitos valores iguais a zero nas entradas posicionadas fora da diagonal principal e um valor relativamente alto e próximo a 1 no elemento da diagonal. Se o documento não compartilha quaisquer termos com o restante dos documentos, então o elemento correspondente à diagonal da matriz $C$ possuirá o valor máximo, isto é 1, e todas as entradas além da diagonal principal terão, para linha equivalente ao documento, valores iguais a zero. As propriedades da matriz $C$ apontam além disso que as entradas $c_{ij}$ podem ser interpretadas como: a extensão com a qual o documento $d_i$ é envolvido pelo documento $d_j$, para $i \neq j$ (acoplamento de $d_i$ com $d_j$); a extensão com a qual o documento $d_i$ é envolvido pelo próprio documento, para $i = j$ (desacoplamento de $d_i$ para com os demais documentos) [7].

Conforme observado a partir das discussões anteriores, se $d_i(1 \leq i \leq m)$ possui poucos termos em comum com os outros documentos, então $c_{ii}$ apresentará um valor mais elevado. A medida expressa por $c_{ii}$ é denominada de coeficiente de desacoplamento, $\delta_i$, de $d_i$, e representa a proporção do quanto um documento distingue-se dos demais documentos. A soma das entradas além da diagonal da $i$-ésima linha da matriz $C$, indica a extensão do acoplamento de $d_i$ com os de-

mais documentos da coleção, sendo referida como coeficiente de acoplamento, $\psi_i = 1 - \delta_i$, de $d_i$. Os valores admitidos por $\delta_i$ e $\psi_i$, situam-se nos intervalos $0 < \delta_i \leq 1$ e $0 \leq \psi_i < 1$, e são base para determinação dos coeficientes médios de desacoplamento, $\delta$, e acoplamento, $\psi$, dos documentos, calculados por meio das expressões $\delta = \sum_{i=1}^{m} \frac{\delta_i}{m},\ 0 < \delta \leq 1,$ e e $\psi = \sum_{i=1}^{m} \frac{\psi_i}{m},\ 0 \leq \psi < 1$ [7].

A matriz $C$, correspondente à matriz de documentos $\times$ termos, $D$, cujas entradas $c_{ij}$ são determinadas de acordo com a expressão $c_{ij} = \alpha_i \sum_{k=1}^{n} d_{ik}\beta_k d_{jk}, 1 \leq i, j \leq m$, encontra-se apresentada a seguir

$$C = \begin{bmatrix} 0,417 & 0,417 & 0,000 & 0,083 & 0,083 \\ 0,313 & 0,438 & 0,000 & 0,063 & 0,188 \\ 0,000 & 0,000 & 0,333 & 0,333 & 0,333 \\ 0,083 & 0,083 & 0,111 & 0,361 & 0,361 \\ 0,063 & 0,188 & 0,083 & 0,271 & 0,396 \end{bmatrix}.$$

Para esta matriz os coeficientes médios de acoplamento e desacoplamento são, nesta ordem, $\delta = \sum_{i=1}^{5} \frac{\delta_i}{5} = \frac{1,945}{5} = 0,389$ e $\psi = 1 - \delta = 0,611$ [7].

No contexto da análise de agrupamentos, a determinação do número de grupos naturalmente apresentado pela coleção de objetos, tem sido considerado um problema irresoluto. Tendo como exemplo os métodos de agrupamento hierárquico, verifica-se que a quantidade de grupos de um conjunto de dados pode variar entre 1 (a coleção completa) e o número de documentos (número de objetos da coleção), e que o dendograma resultante da execução destes algoritmos pode ser segmentado em um dos seus níveis a fim de se obter um número de grupos predeterminado. A operação de segmentação, a qual consiste no critério de parada empregado pelos métodos hierárquicos, é via de regra difícil de ser aplicada, e mesmo para pequenas coleções de dados representa uma condição não facilmente estabelecida. Para uma coleção de documentos arbitrária, o valor extremo inferior para o número de grupos ocorre quando todos os documentos são idênticos, circunstância na qual somente um grupo é observado. O outro extremo para o número de grupos pode ser verificado quando todos os objetos são distintos e a quantidade de grupos é igual ao número de documentos. Em geral, no entanto, o número de grupos será maior do que 1 e menor do que a quantidade de documentos da coleção $(m)$, haja vista que os documentos não serão completamente idênticos e nem completamente distintos. A partir destes fatos e observações torna-se possível estabelecer a hipótese de que o número de grupos de uma coleção de documentos deve ser alto se os documentos são dissimilares, e deve ser baixo caso contrário [7].

De acordo com [7], embora a hipótese seja óbvia, o conceito de similaridade não é de muita utilidade na obtenção do número de grupos, e esta circunstância é resultante da dificuldade em se estipular um limiar de semelhança (critério de parada) que resultará no número de grupos desejado. Em um esforço para fornecer uma solução para este problema, estabelece-se a aplicação do conceito do coeficiente de cobertura na determinação do número de grupos, $n_g$, de uma coleção de documentos, a partir dos elementos posicionados na diagonal da matriz $C$ $n_g = \sum_{i=1}^{m} \delta_i = \delta m$. Observando-se as entradas expressas pela matriz $C$, o número de grupos apresentado pela coleção de documentos retratada por meio da matriz $D$ será determinado por $n_g = \sum_{i=1}^{5} \delta_i = (0,417 + 0,438 + 0,333 + 0,361 + 0,396) = 1,945$ ou $n_g = \delta m = 0,389 \cdot 5 = 1,945 \cong 2$.

Segundo [7], o algoritmo C$^3$M representa uma metodologia de agrupamento de documentos orientada a centróides, por meio da qual $n_g$ documentos são selecionados como centróides dos grupos, e $m - n_g$ documentos não escolhidos como centróides são reunidos em torno dos centróides inicialmente estabelecidos, no intuito de se constituir os grupos pretendidos. Os centróides devem estar bem afastados uns dos outros ao mesmo tempo em que devem ser capazes de associar a si mesmos os demais documentos. Desta forma, os objetos classificados como centróides não devem ser nem muito gerais (contendo numerosos termos) e nem muito específicos (contendo somente alguns termos), e a fim de alcançar este propósito estabelece-se o conceito de aptidão para centróide de grupo, $P_i$, de $d_i (1 \leq i \leq m)$, cujo valor é determinado por $P_i = \delta_i \psi_i \sum_{j=1}^{n} d_{ij}$. Nesta expressão $\delta_i$ representa a separação entre os grupos (dispersão intra-grupos), $\psi_i$ denota a relação entre os documentos que pertencem a um grupo, e o terceiro termo (somatório) fornece a normalização. Por este princípio, os primeiros $n_g$ documentos, organizados em ordem decrescente de $P_i$, deverão ser selecionados como centróides da coleção.

Sob a perspectiva do valor da aptidão verifica-se que os centróides podem ser quase idênticos, haja vista que alguns documentos são eventualmente descritos por um conjunto de termos muito semelhante. Com o propósito de eliminar os centróides aproximadamente iguais, o seguinte método é introduzido: ordenam-se os documentos de acordo com o valor de aptidão para centróide de grupos; a partir das propriedades da matriz $C$, verifica-se se os candidatos sucessivos a centróide, $d_i$ e $d_j$, são praticamente idênticos por meio da análise das entradas $c_{ii}, c_{jj}, c_{ij}$ e $c_{ji}$, pois se os valores absolutos de $(c_{ii} - c_{jj}), (c_{ii} - c_{ij}), (c_{jj} - c_{ji})$ e $(c_{ij} - c_{ji})$ forem inferiores a um dado $\varepsilon$, que equivale ao limiar mínimo de dessemelhança entre os candidatos a centróide, então os documentos $d_i$ e $d_j$ serão qualificados como quase iguais; elimina-se do par constituído pelos candidatos sucessivos e quase idênticos $d_i$ e $d_j$, um dos objetos e considera-se o elemento imediatamente subsequente da lista ordenada de documentos como o novo candidato a centróide [7].

Utilizando como exemplo a matriz $D$ de documentos $\times$ termos, os valores de aptidão para centróide de grupo dos documentos são determinados a partir da expressão $P_i = \delta_i \psi_i \sum_{j=1}^{n} d_{ij}$ e relacionados em ordem decrescente conforme se segue: $P_2 = 0,985, P_5 = 0,957, P_1 = 0,729, P_4 = 0,692$ e $P_3 = 0,222$. Desde que $n_g = 2$, então $d_2$ e $d_5$ tornam-se candidatos a centróides. Tendo em conta que $c_{22} = 0,438$ e $c_{55} = 0,396$, o critério de eliminação com $\varepsilon = 0,001$ estabelece que os objetos são distintos. Desta forma, os documentos $d_2$ e $d_5$ são selecionados como centróides dos grupos

**Algoritmo 6:** Algoritmo C$^3$M

```
Determine os centróides da coleção de documentos;
i ← 1;
enquanto i ≤ m faça
    se d_i não for um centróide então
        Identifique o centróide que maximiza a
        abrangência do conteúdo do documento d_i. Se
        houver mais de um centróide que atenda a esta
        condição, associe d_i ao grupo cujo centróide
        possui, dentre os candidatos, o maior valor de
        aptidão para centróide de grupo;
    i ← i + 1;
Se houver documentos não associados a nenhum dos
grupos, reúna-os em um grupo suplementar (alguns
documentos não classificados como centróides podem
não ter qualquer termo em comum com os objetos
designados como centróides);
```

[7].

De acordo com [7], o método C$^3$M consiste em uma estratégia de agrupamento particional de passagem única, descrita por meio do algoritmo 6.

No intuito de exemplificar o funcionamento do método, considere a construção dos grupos representados pela matriz $D$ de documentos $\times$ termos. Conforme anteriormente determinado, o número de grupos, $n_g$, será 2 e os documentos selecionados como centróides serão $d_2$ e $d_5$. Se $D_o = \{d_1, d_3, d_4\}$ representar a relação dos documentos a serem agrupados e $D_c = \{d_2, d_5\}$ retratar o conjunto dos documentos designados como centróides, então a determinação dos grupos se resumirá ao cálculo dos valores de $c_{ij}$, onde $d_i \in D_o$ e $d_j \in D_c$. Por exemplo, para o documento $d_1$, $c_{12} = 0,417$ e $c_{15} = 0,083$. Desde que $c_{12} > c_{15}$, $d_1$ será associado ao grupo representado pelo centróide $d_2$. Procedendo-se de maneira similar para $d_3$ e $d_4$, obtém-se a partição constituída por $C_1 = \{d_1, d_2\}$ e $C_2 = \{d_3, d_4, d_5\}$ [7].

## 6. Avaliação dos algoritmos de agrupamento

No intuito de realizar a avaliação dos algoritmos de agrupamento pesquisados, dois conjuntos de experimentos foram conduzidos. O primeiro teve como objetivo determinar o melhor método de particionamento iterativo, confrontando entre si as estratégias *K-means*, pesquisa harmônica e algoritmo genético híbrido com a evolução diferencial. Já o segundo grupo de ensaios teve a intenção de comparar o procedimento iterativo mais apropriado com o método C$^3$M, tendo em vista que este é o único algoritmo que adota um modelo de passagem única, não iterativo. Para realizar a avaliação dos procedimentos iterativos, três coleções constituídas de objetos previamente classificados e com número de grupos conhecido foram selecionadas e os elementos constantes das mesmas foram submetidos aos métodos de agrupamento analisados, que foram codificados na linguagem de programação Microsoft Visual Basic .NET e executados em um microcomputador equipado com o sistema operacional Microsoft Windows 7 Professional, memória RAM de 8GB e processador Intel i3 de 2,10GHz.

A primeira coleção foi composta por documentos em idioma inglês, aleatoriamente selecionados dentre os disponibilizados no endereço https://archive.ics.uci.edu/ml/index.html, os quais correspondem aos textos divulgados pela agência de notícias Reuters no ano de 1987. Os documentos obtidos foram subdivididos em três subconjuntos e submetidos às operações de pré-processamento, constituídas das atividades de identificação de termos, lematização das palavras, com o uso do método descrito em [48], e remoção dos termos irrelevantes, a fim de que passassem a ser representados em um formato estruturado, adequado à manipulação por meio dos algoritmos de agrupamento. A segunda coleção foi constituída também por documentos em inglês, casualmente selecionados dentre os disponibilizados no site eletrônico http://qwone.com/ jason/20Newsgroups/, os quais representam 20.000 mensagens eletrônicas classificadas em 20 categorias distintas. De modo análogo ao aplicado para a coleção de textos de notícias Reuters, os documentos da segunda coleção foram igualmente segmentados em três subconjuntos e submetidos às operações de pré-processamento e padronização a fim de que passassem a ser representados de maneira estruturada. A terceira e última coleção, que reunia unicamente documentos em idioma português, foi obtida por intermédio da extração de um subconjunto dos artigos publicados durante setembro de 2015, nos sites eletrônicos dos jornais brasileiros: Correio Braziliense, Diário do Nordeste, O Estado de São Paulo, Folha de São Paulo, Jornal do Brasil, Jornal do Comércio, O Globo e Zero Hora. O critério de seleção dos textos estabelecia que os mesmos deveriam estar presentes nas seções Brasil, Ciência, Cultura, Economia/Negócios, Educação, Espiritualidade/Religião, Esportes, Mundo/Internacional, Política, Saúde, Sociedade ou Tecnologia, de modo a originar uma coleção de objetos constituída por elementos distribuídos em doze grupos. Os documentos desta coleção foram posteriormente subdivididos em três subconjuntos e submetidos às operações de pré-processamento descritas em [6], no intuito de que passassem a ser representados em um formato estruturado, passível de manipulação por intermédio dos algoritmos de agrupamento. As principais características dos conjuntos de textos correspondentes às coleções de objetos empregadas nos experimentos de avaliação, encontram-se descritas na tabela 1 a seguir.

Os métodos iterativos, cujos parâmetros foram estabelecidos conforme descrito na tabela 6, foram comparados sob a perspectiva dos índices de validação de agrupamentos Entropia, Pureza, Rand, Silhueta, Davies-Bouldin e Dunn, descritos em [49, 3, 13, 50, 51], e dos tempos de execução em segundos, adotando-se como medida de dissimilaridade a distância do coseno, referida em [1]. Para cada conjunto de

**Table 1** Características das coleções de objetos textuais utilizadas na avaliação dos métodos iterativos

| Nome | Objetos | Dimensões | Grupos |
|---|---|---|---|
| Reuters 1 | 340 | 1.875 | 4 |
| Reuters 2 | 403 | 1.895 | 8 |
| Reuters 3 | 1.964 | 4.889 | 12 |
| Newsgroups 1 | 200 | 2.159 | 2 |
| Newsgroups 2 | 400 | 3.245 | 4 |
| Newsgroups 3 | 600 | 4.448 | 6 |
| Jornal 1 | 213 | 2.634 | 4 |
| Jornal 2 | 395 | 3.733 | 8 |
| Jornal 3 | 677 | 5.174 | 12 |

textos, que inicialmente foram submetidos às atividades de pré-processamento, os algoritmos foram executados dez vezes, a fim de que as médias dos índices de validação e dos tempos de execução, calculadas após o término do procedimento de agrupamento, pudessem ser comparadas. No intuito de auxiliar a aferição dos resultados, os critérios de avaliação que admitiam resultados além do intervalo compreendido entre 0 e 1 foram normalizados por meio da expressão $x_i^n = \frac{x_i - x_{min}}{x_{max} - x_{min}}$, onde $x_i$ retratava o valor da $i$-ésima ocorrência do resultado, $x_i^n$ correspondia ao valor normalizado da $i$-ésima ocorrência, $x_{min}$ denotava o menor valor observado e $x_{max}$ o maior valor verificado, ou por meio da expressão $x_i^n = \frac{x_{max} - x_i}{x_{max} - x_{min}}$ conforme as melhores respostas fossem, respectivamente, representadas pela maximização ou minimização do critério em análise. Uma tabela de escores, que atribuía valor 1 ao melhor resultado e também 1 a qualquer valor distinto deste, desde que estivesse 5% além ou aquém do mesmo, foi elaborada com o propósito de retratar os achados dos experimentos. Os escores alcançados pelos métodos de particionamento foram somados, sendo o parecer mais favorável atribuído ao algoritmo que obtivesse a maior pontuação.

A tabela 5 expressa os escores obtidos por cada método iterativo quando do particionamento dos objetos presentes nas coleções de objetos referidas na tabela 1. Os valores constantes da tabela indicam uma performance superior do método *K-means* em relação à pesquisa harmônica e ao algoritmo genético híbrido com a evolução diferencial, sugerindo desta forma que o primeiro seria o algoritmo de agrupamento de textos mais adequado, dentre os métodos iterativos avaliados. Em particular, observa-se que o método *K-means* apresenta melhores resultados em cinco dos sete critérios considerados, haja vista que somente para o tempo de execução e para o índice de validação Davies-Bouldin o comportamento deste algoritmo foi menos apropriado. Além disso, verifica-se que os escores obtidos tanto pelo *K-means* quanto pelos demais métodos de particionamento não foram influenciados pelo idioma, tendo em conta que para os textos em inglês e para os textos em português os algoritmos obtiveram resultados equivalentes.

O segundo grupo de ensaios teve como intenção confrontar o melhor procedimento iterativo de particionamento até então identificado, ou seja o *K-means*, com o algoritmo de passagem única C$^3$M, e para este fim duas coleções de objetos não classificados e com número de grupos desconhecido foram utilizadas. A primeira coleção foi constituída por documentos aleatoriamente selecionados do site eletrônico https://archive.ics.uci.edu/ml/index.html, o qual disponibiliza, dentre outros arquivos, uma coletânea de 129.000 resumos de trabalhos premiados pela *National Science Foundantion*, sobre os quais foram determinados três subconjuntos de dados compostos de 100, 300 e 500 documentos, respectivamente denominados como NSF 1, NSF 2 e NSF 3. A segunda coleção de textos, originada do banco de dados de um sistema de gestão de informações jurídicas, foi composta de três subconjuntos de atos processuais provenientes de diários de justiça eletrônicos, nomeados como Atos Processuais 1, Atos processuais 2 e Atos processuais 3, contendo, respectivamente, 100, 300 e 500 textos. As principais características dos conjuntos de textos correspondentes às coleções de objetos empregadas nos experimentos de avaliação, encontram-se descritas na tabela 2 a seguir.

**Table 2** Características das coleções de objetos textuais utilizadas na avaliação dos métodos *K-means* e C$^3$M

| Nome | Objetos | Dimensões | Grupos |
|---|---|---|---|
| NFS 1 | 100 | 857 | 19 |
| NFS 2 | 300 | 1.749 | 33 |
| NFS 3 | 500 | 2.267 | 42 |
| Atos processuais 1 | 100 | 1.083 | 13 |
| Atos processuais 2 | 300 | 2.679 | 24 |
| Atos processuais 3 | 500 | 3.061 | 36 |

Antes que os métodos de particionamento fossem aplicados, as duas coleções foram submetidas às operações de pré-processamento representadas pelas atividades de seleção de termos, remoção de palavras irrelevantes e lematização dos termos, com o emprego do método descrito em [48] para os textos em idioma inglês da coleção NSF, e do algoritmo apresentado em [6] para os documentos em português originados da coleção de atos processuais. Os números de grupos de cada conjunto de dados, referidos na tabela 2 , foram estimados pelo método C$^3$M quando de sua execução, e foram adotados como um dos parâmetros de entrada do algoritmo *K-means*, que permaneceu com os demais critérios inalterados. O algoritmo C$^3$M admitia como critério de entrada somente o limiar de dessemelhança entre os candidatos a centróide, o qual foi, conforme o trabalho de [7], estabelecido em 0,001. A inexistência de uma categorização prévia das coleções de textos empregadas nestes experimentos, impossibilitou a utilização de índices externos de validação de agrupamentos. Desta forma, os índices de validação interna Silhueta, Davies-Boludin e Dunn, acrescidos do tempo de execução, representaram os critérios admitidos na comparação entre os algoritmos. A medida de dissimilaridade utilizada foi a distância do coseno, ressaltando-se que para o algoritmo C$^3$M a aplicabilidade da mesma restringia-se ao cálculo dos

índices de validação interna. De modo semelhante ao empreendido na avaliação dos métodos iterativos, os algoritmos *K-means* e C$^3$M, foram, para cada conjunto de dados, executados dez vezes e as médias dos resultados apresentados por cada critério de análise foram confrontadas por intermédio da tabela de escores 7.

Os valores registrados na tabela 7 assinalam um desempenho superior do *K-means* quando comparado ao C$^3$M, sugerindo deste modo que o método iterativo *K-means* seria o algoritmo de agrupamento de documentos mais adequado dentre os avaliados neste trabalho. Destaca-se ainda que apesar do método C$^3$M exibir resultados invariavelmente piores para os índices de validação de agrupamentos, este algoritmo apresentou um tempo de execução repetidamente melhor, o qual pode ser justificado pela natureza não iterativa do método. Ademais, o fato do algoritmo C$^3$M compreender um expediente capaz de determinar o número de grupos de uma coleção de documentos, o distingue como um método supostamente relevante, haja vista a importância que o número de grupos desempenha em problemas de particionamento de objetos.

Com o objetivo de avaliar o número de grupos estabelecido pelo método C$^3$M ao processar uma coleção de documentos, três conjuntos de dados textuais rotulados, ou seja, com número de grupos previamente conhecido, foram submetidos ao algoritmo proposto por [7]. Os dois primeiros conjuntos de dados, denominados respectivamente de Reuters 5 e Jornal 6, correspondiam à subcoleções dos conjuntos Reuters 2 e Jornal 2, descritos na tabela 1. O conjunto Reuters 5 compreendia 250 objetos, representados por vetores de 1.619 dimensões, previamente classificados em 5 grupos, ao passo que o conjunto Jornal 6 encerrava 100 documentos, retratados por vetores de 1.397 dimensões, antecipadamente categorizados em 8 grupos. A terceira coleção, denominada Artigos 1, foi constituída por 19 artigos previamente classificados em três categorias, publicados em janeiro de 2014 e extraídos do site eletrônico de notícias http://www.opovo.com.br, que após operações de pré-processamento passaram a ser representados por vetores de 449 dimensões. A tabela 3 retrata, além do número de objetos e de grupos presente em cada coleção de documentos, a quantidade de grupos estimada pelo método C$^3$M.

**Table 3** Número de grupos estimado pelo método C$^3$M para coleções de textos rotuladas

| Nome | Objetos | Grupos | Estimado pelo C$^3$M |
|---|---|---|---|
| Reuters 5 | 250 | 5 | 23 |
| Jornal 6 | 100 | 8 | 19 |
| Artigos 1 | 19 | 3 | 6 |

Por intermédio da análise dos achados apresentados na tabela 3, verifica-se que o algoritmo C$^3$M não foi eficiente em determinar de maneira nem mesmo aproximada, os números de grupos presentes em nenhuma das coleções de documentos avaliadas. Estes resultados sugerem que o método não seria adequado à determinação do número de grupos presente nas coleções de atos processuais que se pretende categorizar, assinalando por consequência que estudos complementares, no intuito de indicar um expediente mais preciso, devam ser realizados. Com efeito, a partir de investigações adicionais da literatura pertinente, propõem-se que duas estratégias sejam de modo suplementar avaliadas.

A primeira, estabelecida por [9], consiste no algoritmo polinomial IGN (Identificador de Grupos Naturais), que é baseado em técnicas hierárquicas e que mostra-se eficiente em identificar soluções exatas para diversas instâncias do problema de agrupamento, disponíveis na literatura. Este método, que recebe como entrada o conjunto de $n$ indivíduos e a quantidade mínima de elementos que devem existir em cada grupo, compreende as seguintes etapas:

- Executar um algoritmo de árvore geradora mínima (a exemplo do Prim, do Kruskal ou outro) e armazenar o custo das arestas estabelecidas pela árvore originada;
- Definir uma função de avaliação $F(k) : \mathbb{R} \to \mathbb{R}$, $\forall\, k = 1, \ldots, n-1$ na qual o resultado de $F$ na iteração de ordem $k$ é determinado pela norma da diferença entre o valor da função $f$ para as iterações de ordem $k+1$ e $k-1$, ou seja, $F(k) = \| f(k+1) - f(k-1) \|$, onde: uma iteração representa a remoção de uma aresta da árvore geradora mínima, partindo-se da maior para menor; e $f(k) : \mathbb{R}^n \to \mathbb{R}$ consiste em uma função calculada pela diferença entre o custo da floresta formada em relação ao número $n$ de elementos do conjunto e o número $K$ de árvores (grupos) estabelecido;
- Selecionar a menor diferença entre os valores da função $F$ para duas iterações subsequentes, $min_{2 \leq k \leq n}\{F(k) - F(k-1)\}$;
- Retornar a floresta de melhor resultado levando em consideração a restrição que estipula o número mínimo de indivíduos que devem existir em cada grupo.

Uma segunda estratégia, igualmente apta a determinar automaticamente o número de grupos presente em uma coleção de objetos, é representada pelo algoritmo CLUES, definido por [8] e constituído por três procedimentos essenciais:

- Contração: neste procedimento, cada elemento pertencente à coleção de objetos é interpretado como uma partícula de massa igual a unidade, com velocidade inicial zero e sob a influência de um campo gravitacional que a impulsiona na direção das regiões de maior densidade do conjunto de dados. Por meio deste mecanismo, os elementos do conjunto de dados deverão convergir para os então denominados pontos focais, os quais representarão os centros dos grupos. A contração de cada objeto do conjunto $\{y_i\}, i = 1, \ldots, n$, na direção das regiões de densidade mais elevada é realizada por intermédio de uma estratégia que, de modo iterativo e até que uma condição de convergência seja satisfeita, ajusta o valor de cada coordenada de $y_i$ com base

na mediana dos $K$ vizinhos mais próximos, ou seja, $y_i^{(t+1)} = mediana_{y_j^{(t)} \in N_K(y_i^{(t)})} y_j^{(t)}$, onde $N_K(y^{(t)})$ representa a menor esfera que compreende os $K$ elementos mais adjacentes ao objeto $y_i^{(t)}$. Se $Y$ representar uma matriz $n \times m$ dos elementos do conjunto $\{y_i\}, i = 1, \ldots, n$ no espaço $m$-dimensional $\mathbb{R}^m$ e $y_{ij}$ retratar a $j$-ésima coordenada do $i$-ésimo elemento, então para quaisquer valores fixos $K$, $\varepsilon$ e $M$, o procedimento de contração pode ser descrito conforme a seguir:

1. Inicialize o número da iteração $t \leftarrow 1$ e realize uma cópia dos dados originais, atribuindo $Y^{(t)} \leftarrow Y$ e $Y^{(t+1)} \leftarrow Y$;
2. Atualize as coordenadas de cada objeto $y_i^{(t)}$, por $y_i^{(t+1)} = mediana_{y_j^{(t)} \in N_K(y_i^{(t)})} y_j^{(t)}$;
3. Determine $d \leftarrow max_{1 \leq i \leq n} max_{1 \leq j \leq m} \mid y_{ij}^{(t)} - y_{ij}^{(t+1)} \mid$;
4. Se $d < \varepsilon$ ou $t > M$, então execute a etapa 5, caso contrário, $y_{ij}^{(t)} \leftarrow y_{ij}^{(t+1)}$, $t \leftarrow t+1$, e, em seguida, execute a etapa 2;
5. Retorne $y^{(t+1)}, i = 1, \ldots, n$.

O parâmetro $\varepsilon$, que recebe valores da ordem de $10^{-4}$, representa o critério de interrupção do procedimento, e desde que seja razoavelmente pequeno o algoritmo obterá resultados satisfatórios. O parâmetro $M$ consiste no número máximo de iterações e deve ser definido a fim de prevenir a ocorrência de *loops* infinitos.

- Particionamento: tem como objetivo estabelecer a pertinência dos objetos em relação aos grupos. Neste procedimento, denota-se por $\{y_i^{'}\}, i = 1, \ldots, n$ o conjunto dos elementos resultantes da operação de contração e definem-se dois conjuntos $S_U$ e $S_R$, que compreendem, respectivamente, os rótulos dos elementos já utilizados e os rótulos dos elementos ainda não utilizados pelo método. Inicialmente, seleciona-se um elemento arbritário de $S_R$, denotado por $y_a^{'}$, e então procura-se, a partir do emprego de uma função $d$ que estabelece a distância entre os objetos, obter o elemento mais próximo de $y_a^{'}$, nomeado como $y_b^{'}$. A seguir, atribui-se $y_a^{'}$ a $S_U$, substitui-se $y_a^{'}$ por $y_b^{'}$ e então repete-se o procedimento de identificação do elemento mais adjacente. A estratégia de particionamento é baseada na ordem em que os elementos são incluídos no conjunto $S_U$ e na distância verificada entre dois elementos subsequentes, sugerindo-se que intervalos significativos indicam o início de um novo grupo. No intuito de evitar interpretações subjetivas quanto ao que determinaria um intervalo significativo, o método estabelece que intervalos desta natureza serão os que estiverem a $1,5 \cdot IIQ$ de distância da média de todas as distâncias, onde $IIQ$ corresponde ao intervalo inter-quartil das distâncias entre os elementos de $S_U$. Uma vez que os intervalos significativos tenham sido identificados, quaisquer objetos posicionados entre dois destes intervalos supostamente pertencerão ao mesmo grupo. As etapas executadas pelo algoritmo de particionamento, que nesta circunstância emprega a distância Euclidiana, mas que admite o uso de outros índices de proximidade, podem ser descritas conforme a seguir:

  1. Inicialize o número da iteração $t \leftarrow 1$, $S_U = \emptyset$ e $S_R = \{1, \ldots, n\}$;
  2. Selecione aleatoriamente um rótulo de $S_R$ e o defina como $a$;
  3. Se $S_R \neq \emptyset$, então
     (a) Identifique $b \in S_R$ tal que $b = min_{1 \leq j \leq n} d(y_a^{'}, y_j^{'})$, ou seja, $y_b^{'}$ é o vizinho mais próximo de $y_a^{'}$;
     (b) $d_t \leftarrow d(y_a^{'}, y_b^{'})$;
     (c) $S_R \leftarrow S_R - \{b\}$ e $S_U \leftarrow S_U + \{b\}$;
     (d) $t \leftarrow t+1$ e $a \leftarrow b$;
     (e) Execute a etapa 3.
  4. Calcule $R = \sum_{i=1}^{n-1} d_i/(n-1) + 1,5 \cdot IIQ$;
  5. Inicialize $C_i, i = 1, \ldots, n$, o número de grupos como $g = 1$ e o número de iterações como $t = 1$;
  6. Se $d_t > R$, então $g \leftarrow g+1$;
  7. $C_{S_U(t+1)} \leftarrow g$, onde $S_U(t+1)$ consiste no $(t+1)$-ésimo elemento do conjunto $S_U$;
  8. Se $t < n$, então $t \leftarrow t+1$, e, em seguida, execute a etapa 6. Caso contrário, execute a etapa 9;
  9. Retorne $C_i, i = 1, \ldots, n$.

- Determinação do número de grupos mais favorável: o agrupamento obtido pelo procedimento de particionamento depende do resultado alcançado pelo procedimento de contração, e a escolha do número de vizinhos mais próximos representa um aspecto fundamental para os dois procedimentos. A fim de evitar que o número de vizinhos tenha que ser especificado como um parâmetro, propõe-se que o algoritmo obtenha o valor de $K$ com base na avaliação da robustez da partição, determinada por intermédio da aplicação de índices de validação de agrupamentos, a exemplo do índice silhueta e do índice de Calinski e Harabasz. No intuito de obter o número ideal de vizinhos $K$, o algoritmo CLUES poderia determinar o resultado do índice de validação selecionado para $K$ de 2 até $n-1$, e, ao final deste procedimento, optar pelo valor de $K$ que originasse o agrupamento mais adequado. Embora válida sob a perspectiva teórica, esta conduta tornaria o método muito demorado, e, com o objetivo de acelerar este processo, adota-se o fator de velocidade $\alpha$, $0 < \alpha < 1$, de modo que o valor inicial de $K$ e o seu incremento sejam iguais a $\alpha n$. Desta forma, se existirem, por exemplo, 1.000

objetos no conjunto de dados e o valor de $\alpha$ for 0,05, então o algoritmo iniciará com $K = 50$, e terá incrementos sucessivos iguais a 50. Por meio deste mecanismo, o algoritmo realizará a exploração em somente uma fração dos objetos do conjunto de dados, conforme determinado pelo fator $\alpha$, e, por consequência, o método convregirá mais rapidamente. O incremento da velocidade de obtenção dos resultados poderá ocasionar um decréscimo na precisão da solução, e, com o objetivo de retificar este problema, um procedimento de refinamento pode ser aplicado uma vez que a estimativa aproximada do valor de $K$ ideal tenha sido obtida. O processo de refinamento corresponde essencialmente ao mesmo procedimento de exploração de soluções, exceto por atribuir a $K$ acréscimos iguais a 1. Por exemplo, suponha que utilizando um fator $\alpha = 0,05$ o método CLUES obtém um $K$ aproximado igual a 250, para uma coleção de 980 objetos e valor de $K$ ótimo de 257. Nessa situação, o algoritmo irá, na etapa de refinamento, avaliar os valores de $K$ entre 201 (250 - $\alpha n$) e 299 (250 + $\alpha n$) no intuito de obter o resultado mais adequado para $K$. Com o propósito de ampliar ainda mais a robustez e a velocidade do método, em identificar o número ideal de objetos mais adjacentes, o procedimento de contração dos pontos para o próximo valor de $K$ é realizado tomando como base as coordenadas modificadas de cada objeto, obtidas quando da contração dos elementos levando em consideração o valor anterior de $K$. Além disso, a influência eventualmente exercida por valores atípicos, ou seja, *outliers*, também é observada. Por conseguinte, com a intenção de impedir a formação de grupos ilegítmos, constítuidos por valores discrepantes, e partindo do princípio de que o número de *outliers* é, em geral, relativamente pequeno, a quantidade mínima de elementos que um grupo deve compreender é estabelecida como $\alpha n$. Se $n$ representar o número de objetos no espaço $m$-dimensional $\mathbb{R}^m$, $g$ o número de grupos, $C$ a partição dos elementos, $T$ o número de mínimo de elementos que um grupo deve encerrar e $\delta$ o incremento aplicado sobre o número de vizinhos mais próximos, então o algoritmo CLUES será definido conforme a seguir:

1. Inicialize $\alpha \leftarrow 0,05$, $K \leftarrow \alpha n$, $\delta \leftarrow \alpha n$, $T \leftarrow \alpha n$, $t \leftarrow 0$, $Y^{(t)}_{n\times m} \leftarrow Y_{n\times m}$ e $s_{max} \leftarrow -100$, onde $s_{max}$ representa o valor máximo de robustez de uma partição;
2. $t \leftarrow t+1$; Realize o procedimento de contração e estabeleça $Y^{(t)}_{n\times m} \leftarrow Y^{(t-1)}_{n\times m}$;
3. Execute o procedimento de particionamento e obtenha $C^{(t)}$ baseado em $Y^{(t)}_{n\times m}$;
4. Se $g^{(t)} > 1$, então
   (a) Obtenha o valor de índice de validação de agrupamentos selecionado $s^{-(t)}$;
   (b) Se $t = 1$, então $s_{max} \leftarrow s^{-(t)}$, $g^* \leftarrow g^{(t)}$ e $C^* \leftarrow C^{(t)}$;
   (c) Se $min_{1\le i\le g} n_i \ge T$ e $s^{-(t)} > s_{max}$, então $s_{max} \leftarrow s^{-(t)}$, $g^* \leftarrow g^{(t)}$, $C^* \leftarrow C^{(t)}$, onde $n_i$ representa o número de objetos do $i$-ésimo grupo;
   (d) Se $min_{1\le i\le g} n_i \ge T$ e $g^* = 2$, então execute a etapa 7;
   (e) Se $min_{1\le i\le g} n_i \ge T$ e $g^{(t)} = 2$, então execute a etapa 7;
   (f) $K \leftarrow K + \delta$; Execute a etapa 2;
5. Caso contrário, se $g^{(t)} = 1$ e $t = 1$, então $g^* \leftarrow 1$, $C^* \leftarrow \{1,\ldots,1\}$ e execute a etapa 7;
6. Caso contrário, se $g^{(t)} = 1$ e $t > 1$, então execute a etapa 7;
7. Retorne o número ótimo de grupos $g^*$ e a partição final $C^*$.

O algoritmo CLUES, que além de estabelecer o número de grupos da coleção determina de que modo os objetos devem ser particionados, admite como parâmetro de entrada somente o critério de convergência $\alpha$. A abordagem baseada nos $K$-vizinhos mais próximos permite que o método possa ajustar-se à estrutura geométrica local do conjunto de dados e que seja capaz de manipular objetos esparsos e de numerosas dimensões [8].

Com o objetivo de avaliar os métodos de particionamento IGN e CLUES, quanto à determinação do número de grupos presente em uma coleção de documentos textuais, os mesmos conjuntos de dados utilizados na apreciação do algoritmo C$^3$M, foram submetidos aos procedimentos IGN e CLUES, que fizeram uso distância do coseno como índice de determinação da dissimilaridade entre os objetos. A tabela 4 demonstra os números de grupos estimados pelos dois métodos

**Table 4** Número de grupos estimado pelos métodos IGN e CLUES para coleções de textos rotuladas

| Nome | Objetos | Grupos | IGN | CLUES |
|---|---|---|---|---|
| Reuters 5 | 250 | 5 | 6 | 23 |
| Jornal 6 | 100 | 8 | 5 | 19 |
| Artigos 1 | 19 | 3 | 3 | 5 |

A análise dos resultados demonstrados pela tabela 4 possibilita verificar que o método de particionamento que de maneira mais aproximada determinou o número de grupos efetivamente presente nas coleções avaliadas foi o IGN, sugerindo, portanto, que este seria, dentre os métodos estudados, o algoritmo mais adequado em estabelecer o número de grupos naturalmente existente em uma coleção de documentos.

## 7. Método de agrupamento e categorização dos atos processuais digitais

### 7.1 Considerações preliminares

De acordo com os resultados apresentados pelos procedimentos de avaliação descritos na seção 6, sugere-se que o método a ser adotado para realizar o agrupamento e a categorização dos atos processuais, deverá: utilizar a medida do coseno como índice de determinação da dissimilaridade entre os textos; empregar uma estratégia que possibilite determinar o número de grupos $K$ que uma coleção de documentos encerra; aplicar o método de particionamento *K-means*, a fim de agrupar os atos processuais nos $K$ conjuntos previamente estabelecidos pelo algoritmo que assinalou o número de grupos. O método proposto deverá ser incorporado ao procedimento do sistema de gestão de informações jurídicas, que diariamente realiza a importação dos atos processuais originados de diários de justiça eletrônicos, com o objetivo de adicionalmente categorizá-los e descrevê-los, por intermédio da atribuição de descritores que deverão caracterizar de forma breve e representativa o seu conteúdo. As etapas que deverão ser efetuadas pelo sistema de gestão de informações quando do processamento das publicações jurídicas, ao considerar-se os procedimentos até então executados, e os novos, impostos pela introdução do método de agrupamento de documentos, serão:

- Importar para o banco de dados do sistema os atos processuais digitais, por meio da utilização dos serviços de uma empresa especializada em captura de publicações divulgadas pelos diários de justiça eletrônicos;
- Executar o tratamento e a padronização dos textos escritos em linguagem natural com o uso dos procedimentos de eliminação de variações morfológicas e remoção de termos irrelevantes, conforme orientações descritas em [6], a fim de que as publicações passem a ser representadas em um formato estruturado, passível de manipulação por parte dos algoritmos de agrupamento de textos. O modelo estruturado adotado será o espaço vetorial, no qual cada documento é representado como um *bag-of-words* que utiliza as palavras como medida para identificar a similaridade entre os objetos. Por intermédio deste esquema, cada documento $d_i$ será considerado um ponto no espaço vetorial $m$-dimensional, no qual a dimensão $m$ corresponderá ao número de termos distintos da coleção de documentos e o valor de cada componente de $d_i$ será determinado conforme a expressão *tf-idf*;
- Obter o número de grupos $K$ da coleção de atos processuais, utilizando a estratégia definida pelo algoritmo IGN;
- Realizar o particionamento dos textos em $K$ grupos com o emprego do algoritmo *K-means*;
- Atribuir descritores aos grupos determinados pelo algoritmo *K-means*, por intermédio da seleção, para cada grupo, dos cinco termos mais representativos, de acordo com o critério indicado pelo coeficiente de correlação, conforme as orientações descritas em [17].

Os descritores conferidos aos atos processuais serão representados pelos termos atribuídos como descritores dos grupos aos quais os mesmos estarão associados. Desta forma, uma publicação jurídica demonstrada pelo sistema de gestão de informações deverá apresentar, além do texto em si, uma relação de termos que terão a intenção de caracterizá-la de forma breve e significativa, facilitando a análise do seu teor por parte dos interessados.

No intuito de exemplificar o comportamento do método proposto, uma coleção constituída por 100 atos processuais, $A = \{a_i\}, i = 1, \ldots, 100$, selecionados dentre as publicações importadas no período compreendido entre 01/01/2015 e 31/03/2015, foi submetida ao processo de agrupamento e categorização de textos. Inicialmente, os documentos foram convertidos em um formato estruturado, passando a ser representados por meio de vetores de 1.458 dimensões, obtidos por intermédio da aplicação dos procedimentos de tratamento e padronização de textos. Em seguida, e no intuito de determinar o número de grupos $K$ presente na coleção, o algoritmo IGN foi executado, auferindo como resultado 4 grupos. Por fim, com o objetivo de categorizar os atos processuais, a coleção foi organizada pelo algoritmo de particionamento *K-means*, que admitiu como um dos parâmetros de entrada $K = 4$ e que atribuiu como descritores de cada grupo os cinco termos mais representativos, selecionados de acordo com o raciocínio estabelecido pelo coeficiente de correlação. A distribuição dos atos entre os grupos, acrescidos dos respectivos descritores, conforme determinada pelo algoritmo *K-means* quando do emprego da representação *bag-of-words*, é demonstrada pela tabela 8.

A tabela 9 ilustra a avaliação do algoritmo *K-means* em termos do tempo de execução e dos índices Silhueta, Davies Bouldin e Dunn, quando aplicado na determinação da partição com o uso da representação *bag-of-words*.

Por intermédio de duas amostras obtidas da partição determinada pelo *K-means* verificou-se que:

- Os descritores do ato processual $a_{46}$, com redação igual a "DESPACHOS-3ª Câmara Cível Serviço de Recursos da 3ª Câmara DECISÃO MONOCRÁTICA - Nº 0188860-23.2012.8.06.0001-Apelante: Maritima Seguros S/A-Apelado: Francisco Ferreira do Nascimento-Diante do exposto, dou provimento à apelação, para reformar a sentença proferida pelo Juízo da 20ª Vara Cível da Comarca de Fortaleza, para julgar totalmente improcedente o pedido exordial, o que faço com fundamento no art.557, §1º - A, do Código de Processo Civil, considerando a jurisprudência dominante e de efeito vinculante do Supremo Tribunal Federal. Inverto, ainda, os ônus sucumbenciais, condenando o apelado no pagamento de R$ 1.000,00 (mil reais) a título de honorários advocatícios. Por ser o autor/apelado beneficiário da justiça gratuita, fica suspensa a exigibilidade pelo prazo de cinco (05) anos ou até a comprovação superveniente da cessação do estado de pobreza, conforme previsto no art.12, da Lei 1060/50. Fortaleza, 18 de dezem-

bro de 2014. DESEMBARGADOR WASHINGTON LUIS BEZERRA DE ARAUJO Relator - Advs: Carlos Robson Nogueira Lima Filho (OAB: 21231/CE)-Rostand Inacio dos Santos (OAB: 22718/PE)-Leonardo Araujo de Souza (OAB: 15280/ CE) - 1124 - Diário da Justiça do Ceará - LEONARDO ARAUJO DE SOUZA - 15280" e pertencente ao grupo 1, foram os termos "cível", "vara", "oliveira", "secretaria" e "danielle";

- Os descritores do ato processual $a_{90}$, também pertencente ao grupo 1 e de conteúdo igual a "Notificação - Processo Nº RTOrd-0001739-16.2013.5.07.0007 Relator Francisco Antonio da Silva Fortuna Reclamante Lidiane de Menezes de Sousa Advogado Pedro Paulo Silva de Oliveira (OAB: 23929) Reclamado Jose Airton Farias de Oliveira Advogado Gerardo Majela de Castro (OAB: 11812- B) Reclamado Tania Maria Farias de Oliveira Advogado Raimundo Augusto Fernandes Neto (OAB: 6615) Advogado Esio Rios Lousada Neto (OAB: 18190) Fica o advogado Pedro Paulo Silva de Oliveira, notificado para ciência da expedição de alvará de ID 4c7c8c1. Notificação realizada via DEJT conforme Resolução CSJT Nº136/2014 - 1660 - Diário da Justiça do Ceará - RAIMUNDO AUGUSTO FERNANDES NETO - 6615", foram os mesmos termos associados ao ato processual $a_{46}$, ou seja, "cível", "vara", "oliveira", "secretaria" e "danielle".

Uma análise realizada por profissionais da área de advocacia, sobre os agrupamentos de atos processuais e seus descritores, permitiu concluir que embora os elementos que caracterizavam os grupos de textos tivessem sido determinados a partir de procedimentos que avaliaram a significância estatística dos termos, o resultado obtido não teve a relevância esperada originalmente. A despeito das palavras constituintes dos descritores estarem, em geral, presentes nos atos processuais, o valor semântico das mesmas em revelar de modo sucinto e significativo o teor dos documentos encerrados por um dado grupo não foi expressivo. Com efeito, de acordo com [52], [53], [54] e [17], entre outros, uma das maiores dificuldades pertinentes ao processo de identificação e análise de grupos de documentos, para qualquer que seja o algoritmo de particionamento adotado, consiste na seleção inapropriada das características que descrevem e modelam os textos. Os elementos utilizados para interpretar o conteúdo dos documentos são em geral representados pelas palavras neles compreendidas, e, apesar do emprego de estratégias que procuram identificar os termos mais relevantes, uma escolha fundamentada somente em critérios quantitativos resulta na seleção de palavras desprovidas de contexto, que não representam adequadamente o conteúdo de um documento. Como emprego de palavras isoladas e desconexas prejudica a identificação das ideias, objetos ou ações presentes nos textos, o usuário passa a necessitar de um conhecimento prévio acerca dos documenou dos assuntos retratados pelos mesmos, a fim de melhor compreendê-los por meio da observação dos descritores dos grupos. Por consequência, a análise exploratória dos textos fica comprometida, haja vista que a subjetividade na interpretação dos resultados torna-se ainda maior.

Cumpre ressaltar que uma abordagem alternativa à seleção dos descritores via coeficiente de correlação que poderia ser adicionalmente avaliada, seria a modelagem LDA (*Latent Dirichlet Allocation*), que é referida e empregada em [55, 56, 57, 58] e que é fundamentada na concepção de que cada documento é representado como uma combinação de tópicos e de que cada tópico é representado por um conjunto de termos (unigramas, bigramas ou $n$-gramas), ambos com probabilidades associadas que indicam a maior ou menor relevância do termo em relação ao tópico assim como do tópico em relação ao documento. Entretanto e em consequência de suas características fundamentalmente quantitativas, esta estratégia não foi suplementarmente examinada.

### 7.2 Utilização de conceitos na representação de documentos

A fim de procurar suplantar os problemas resultantes da seleção inadequada dos descritores, sugere-se, conforme o trabalho de [52], a substituição das palavras utilizadas para representar os documentos por estruturas mais apropriadas, aptas a modelar as ideias presentes nos textos e a facilitar a sua compreensão. Para este fim, tais estruturas devem ser capazes de aproximar o conteúdo do documento ao usuário, por intermédio da utilização de um vocabulário que lhe seja habitual. Uma maneira de estabelecer uma relação entre o conteúdo de um documento e o vocabulário do usuário consiste em representar os textos por intermédio de conceitos. Os conceitos correspondem a fragmentos que o ser humano utiliza para representar ideias, opniões ou pensamentos e que podem ser expressos por meio de termos específicos, ou seja, palavras que quando encontradas assinalam a presença do conceito com um determinado nível de relevância. Desde que se encontrem corretamente modelados e definidos de maneira a representar um conhecimento que esteja no contexto do interessado, a utilização dos conceitos permite retratar o conteúdo do documento em um nível de abstração mais elevado, auxiliando o usuário a compreender melhor os resultados originados pelos métodos de determinação e análise de agrupamentos [52].

O propósito primordial e imprescindível da modelagem por conceitos, resume-se em considerar os conceitos presentes nos documentos durante o procedimento de pré-processamento que estabelece as características representativas dos textos. O mecanismo de pré-processamento recebe a coleção de documentos, determina os conceitos pertinentes e origina uma estrutura de representação intermediária, que denota cada documento por meio dos conceitos identificados e que é utilizada pelo algoritmo de particionamento, durante o processo que fará a categorização dos textos entre os grupos [52]. De acordo com [52], um conceito é composto por um identificador e por um conjunto de palavras que o descrevem. O identificador representa a ideia geral do conceito e pode corresponder a nomes de objetos, substantivos, ações ou qualquer

outro termo que tenha relevância para o usuário. O conjunto de descritores do conceito compreende palavras que assinalam a presença do conceito no documento e que podem ser representadas pelo próprio identificador juntamente com outras palavras relacionadas, incluindo-se variações morfológicas ou ainda erros ortográficos. Por este modelo, o conceito ”futebol” poderia, por exemplo, ter como identificador o próprio termo ”futebol” e possuir como descritores as palavras ”bola”, ”trave”, ”campo”, ”pênalti”, ”jogador”, ”jogo”, ”juiz” e ”escanteio”. De modo similar, o conceito ”cancelar” poderia ser identificado pela palavra ”cancelar” e descrito pelos termos ”cancela”, ”cancelar”, ”cancelarei” e ”cancel”.

No estudo realizado por [52], os conceitos foram identificados por intermédio de um método baseado no modelo espaço vetorial, no qual cada conceito é representado por um vetor de termos que admite a existência de níveis distintos de relação entre os termos e os conceitos, de modo que termos com graus de relação mais elevado indicam a presença imediata do conceito, ao passo que termos com níveis de relação mais reduzidos exigem a identificação de outros termos para que a existência do conceito no documento possa ser assinalada. Uma função de ativação, que compara os termos que definem o conceito com os termos presentes no documento, foi utilizada para determinar o grau de relacionamento entre os documentos e conceitos. Os valores originados da função de ativação estavam no intervalo $[0,1]$, no qual valores iguais a zero indicavam a ausência do conceito, valores maiores do que zero denotavam a presença do conceito e valores iguais a um determinavam o nível máximo de ativação.

### 7.3 Método baseado em conceitos para agrupamento e categorização dos atos processuais digitais

Neste trabalho, a estratégia que estabelece uma estrutura de conceitos composta por um identificador e por um conjunto de descritores será admitida conforme descrita por [52], porém sugere-se a utilização de um modelo de representação vetorial distinto do originalmente proposto, elaborado com base nos estudos de [59, 60, 61] e definido de maneira a retratar cada texto como um *bag-of-concepts*. Por este princípio, cada documento $d_i$ será um ponto no espaço vetorial $m$-dimensional, $d_i = (w_{i1}, w_{i2}, \ldots, w_{im}), i = 1, \ldots, n$, no qual a $m$-ésima dimensão corresponderá ao número de conceitos distintos da coleção de documentos. Cada componente de $d_i$ representará um conceito da coleção que pode ou não estar presente no documento, e o valor de cada componente dependerá do grau de relacionamento entre o conceito e o documento que possivelmente o contém, calculado pela expressão $w_{ij} = n_{ij} log_2 \left(\frac{n}{n_j}\right)$, onde $n_{ij}$ denota a frequência do conceito, ou seja quantas vezes os descritores do conceito $c_j$ ocorrem no documento $d_i$, $n_j$ corresponde ao número de documentos nos quais o conceito $c_j$ sucede e $n$ representa o número de documentos da coleção.

Com a finalidade de ilustrar o comportamento do método proposto, quando da substituição do modelo de representação vetorial *bag-of-words*, pelo modelo *bag-of-concepts*, a mesma coleção de documentos a princípio utilizada, constituída por 100 atos processuais importados no período compreendido entre 01/03/2015 e 31/03/2015, foi submetida ao processo de agrupamento e categorização de textos. Inicialmente e com base nas necessidades e experiências anteriores quanto à leitura de atos processuais, de um escritório de advocacia especializado em ações revisionais de contratos, um conjunto composto por seis conceitos foi definido, em seguida, os atos processuais foram convertidos em um formato estruturado, e passaram a ser representados por intermédio de vetores determinados com a aplicação do modelo *bag-of-concepts*. Posteriormente, e no intuito de estabelecer o número de grupos da coleção de atos processuais, o algoritmo IGN foi executado, subsequentemente, os atos processuais foram particionados em grupos com o emprego do método *K-means*, por fim, os identificadores dos conceitos mais representativos, determinados pelo coeficiente de correlação e limitados até cinco, foram atribuídos como descritores dos grupos.

Os identificadores e os descritores dos conceitos, indicados pelos responsáveis que analisavam os atos processuais e instituídos levando em consideração a relevância que dados elementos presentes nos textos exerciam na caracterização do seu conteúdo, encontram-se descritos na tabela 10.

Na abordagem exemplificada, o procedimento de pré-processamento foi representado pela substituição do texto do ato processual pelos identificadores dos conceitos, levando em consideração o número de ocorrências dos descritores no conteúdo original do documento. Desta forma, se, por exemplo, o conceito ”Sentença” ocorresse duas vezes em um ato processual, e, de modo similar, se fosse verificado que os descritores do conceito ”Acórdão” estivessem citados três vezes no mesmo ato processual, então o texto primário do ato processual seria permutado pela sequência: ”sentença, sentença, acórdão, acórdão, acórdão”. A partir do texto pré-processado, os documentos passaram ser representados por intermédio de vetores cujo número de dimensões correspondia ao número de conceitos distintos da coleção, isto é, seis, e cujos componentes eram calculados considerando a relação entre os identificadores dos conceitos e os documentos, conforme definido pela expressão $w_{ij} = n_{ij} log_2 \left(\frac{n}{n_j}\right)$. A quantidade de grupos da coleção foi estabelecida pelo algoritmo IGN, com base no processamento da representação conceitual dos atos processuais, e o resultado obtido, $K = 4$, foi utilizado como um dos parâmetros de entrada do método *K-means*, que realizou a categorização dos documentos. A distribuição dos atos entre os grupos, acrescidos dos respectivos descritores quando do emprego da representação *bag-of-concepts*, é demonstrada pela tabela 11.

Por meio de uma nova análise, realizada sobre a amostra constituída pelos atos processuais $a_{46}$ e $a_{90}$, verificou-se que: os descritores do documento $a_{46}$, que inicialmente eram ”cível”, ”vara”, ”oliveira”, ”secretaria” e ”danielle” passaram a ser ”audiência de conciliação”, ”acórdão” e ”sentença”; os descritores do ato $a_{90}$, que anteriormente eram ”cível”, ”vara”,

”oliveira”, ”secretaria” e ”danielle” tornaram-se ”sentença” e ”expedição de alvará”. Uma avaliação complementar dos resultados, efetuada pelos profissionais que possuíam, entre outras atribuições, a responsabilidade de realizar a leitura dos atos processuais, permitiu concluir que o novo formato representou verdadeiramente uma evolução quanto à relevância dos elementos que descreviam os textos. De fato, tendo em vista que pela abordagem *bag-of-concepts*, os descritores dos atos foram originados por meio de conceitos e identificadores estabelecidos pelos próprios interessados, que levaram em consideração o contexto e a importância destes elementos em indicar o conteúdo de um documento, seria plausível pressupor o melhor comportamento apresentado por este método.

A tabela 12 ilustra a avaliação do algoritmo *K-means* em termos do tempo de execução e dos índices Silhueta, Davies Bouldin e Dunn, quando aplicado na determinação da partição com o uso da representação *bag-of-concepts*

Uma comparação entre os valores registrados nas tabelas 9 e 12 sugere um melhor comportamento da representação *bag-of-concepts* diante do modelo *bag-of-words*. Em particular, verifica-se uma redução significativa do tempo de execução do algoritmo de particionamento, proporcionada pelo expressivo descréscimo do número de dimensões dos vetores utilizados na representação estruturada dos textos, associada com uma melhor avaliação em termos de dois, dentre os três índices de validação interna utilizados. Estes resultados atestam a predisposição dos conceitos em representar de maneira breve e significativa o conteúdo dos textos, sem prejuízos em relação à qualidade das partições sob o aspecto estatístico e com uma relevante atenuação da complexidade do processamento necessário à determinação dos grupos. Ressalta-se ainda que o modelo de representação fundamentado em conceitos pode, eventualmente, originar partições que não compreendem todos os objetos presentes na coleção. Na verdade, como a representação vetorial dos textos é dependente da presença dos descritores dos conceitos, alguns objetos podem ocasionalmente ser expressos por estruturas cujos componentes são invariavelmente iguais a zero, em situações nas quais nenhum dos descritores previamente estabelecidos são verificados dentre os elementos que constituem o documento. Tal circunstância inviabiliza o cálculo da dissimilaridade e impede que o texto seja associado a qualquer um dos grupos, haja vista que não é possível estabelecer o quão próximo o objeto se encontra de um determinado centróide. De fato, levando-se em consideração a coleção exemplificada, verifica-se que dentre os 100 atos processuais submetidos ao procedimento de particionamento com o emprego da estratégia *bag-of-concepts*, 17 permaneceram não incorporados aos grupos preliminarmente designados. Este resultado sugere que os identificadores dos seis conceitos estabelecidos não foram suficientemente abrangentes ou que os documentos na verdade pertenciam a um ou mais conceitos não modelados, e denota a influência que a estruturação dos conceitos, por meio de identificadores e descritores relevantes, exerce sobre os resultados alcançados.

Tendo em conta os melhores resultados obtidos pelo modelo *bag-of-concepts* em particionar os documentos e em representá-los de maneira concisa e expressiva, sugere-se que as etapas que deverão ser efetuadas quando do processamento das publicações jurídicas serão:

- Definir, conforme o contexto e a relevância, quais conceitos, acrescidos de seus respectivos identificadores e descritores, serão utilizados para representar os atos processuais;
- Executar o tratamento e a padronização dos textos escritos em linguagem natural, a fim de que as publicações passem a ser representadas em um formato estruturado, passível de manipulação por parte dos algoritmos de agrupamento de textos. O modelo estruturado adotado será o espaço vetorial, no qual cada documento é representado como um *bag-of-concepts* que utiliza os descritores dos conceitos como medida para identificar a similaridade entre os objetos;
- Obter o número $K$ de grupos da coleção de atos processuais, utilizando a estratégia definida pelo algoritmo IGN, e realizar o particionamento dos textos em $K$ grupos com o emprego do algoritmo *K-means*;
- Atribuir descritores aos grupos determinados pelo algoritmo *K-means*, por meio da seleção, para cada grupo, de até cinco identificadores de conceitos levando em consideração os mais representativos, de acordo com o critério determinado pelo coeficiente de correlação, conforme as orientações descritas em [17].

As partições obtidas pelo método proposto serão sobremaneira dependentes da forma como os conceitos serão estruturados, e, para uma mesma coleção de textos, o número de grupos, seus respectivos descritores e até mesmo a quantidade de objetos eventualmente não categorizados, podem ser bem distintos. A determinação do número de conceitos, e a seleção dos seus identificadores e descritores serão subordinadas à natureza da análise que se deseja empreender, e constituem atividades que poderão ser refeitas de modo iterativo no intuito de aperfeiçoar, sob o aspecto qualitativo, os resultados até então estabelecidos.

**Table 5** Análise comparativa dos métodos iterativos. RPH: Resultado pesquisa harmônica; RPHN: Resultado pesquisa harmônica normalizado; EPH: Escore pesquisa harmônica; RAG: Resultado algoritmo genético; RAGN: Resultado algoritmo genético normalizado; EAG: Escore algoritmo genético; RKM: Resultado *K-means*; RKMN: Resultado *K-means* normalizado; EKM: Escore *K-means*

| Critério | Conjunto de dados | RPH | RPHN | EPH | RAG | RAGN | EAG | RKM | RKMN | EKM |
|---|---|---|---|---|---|---|---|---|---|---|
| Tempo de execução em segundos | Reuters 1 | 112 | 1,0000 | 1 | 121 | 0,9369 | 0 | 257 | 0,0000 | 0 |
| | Reuters 2 | 169 | 0,9996 | 1 | 169 | 1,0000 | 1 | 591 | 0,0000 | 0 |
| | Reuters 3 | 10.069 | 1,0000 | 1 | 10.236 | 0,9897 | 1 | 26.249 | 0,0000 | 0 |
| | Newsgroups 1 | 44 | 1,0000 | 1 | 45 | 0,9896 | 1 | 91 | 0,0000 | 0 |
| | Newsgroups 2 | 296 | 0,9846 | 1 | 290 | 1,0000 | 1 | 684 | 0,0000 | 0 |
| | Newsgroups 3 | 841 | 1,0000 | 1 | 855 | 0,9880 | 1 | 2.056 | 0,0000 | 0 |
| | Jornal 1 | 60 | 1,0000 | 1 | 94 | 0,6936 | 0 | 169 | 0,0000 | 0 |
| | Jornal 2 | 316 | 0,9925 | 1 | 310 | 1,0000 | 1 | 1.148 | 0,0000 | 0 |
| | Jornal 3 | 1.243 | 1,0000 | 1 | 1.250 | 0,9981 | 1 | 4.567 | 0,0000 | 0 |
| Entropia | Reuters 1 | 0,7852 | 0,7852 | 0 | 0,6948 | 0,6948 | 0 | 0,3013 | 0,3013 | 1 |
| | Reuters 2 | 0,6394 | 0,6394 | 0 | 0,6131 | 0,6131 | 0 | 0,4729 | 0,4729 | 1 |
| | Reuters 3 | 0,7032 | 0,7032 | 0 | 0,6797 | 0,6797 | 0 | 0,4805 | 0,4805 | 1 |
| | Newsgroups 1 | 0,9926 | 0,9926 | 0 | 0,9523 | 0,9523 | 0 | 0,7350 | 0,7350 | 1 |
| | Newsgroups 2 | 0,9109 | 0,9109 | 0 | 0,9009 | 0,9009 | 0 | 0,7059 | 0,7059 | 1 |
| | Newsgroups 3 | 0,9238 | 0,9238 | 0 | 0,8989 | 0,8989 | 0 | 0,6347 | 0,6347 | 1 |
| | Jornal 1 | 0,8543 | 0,8543 | 0 | 0,8294 | 0,8294 | 0 | 0,7610 | 0,7610 | 1 |
| | Jornal 2 | 0,7757 | 0,7757 | 0 | 0,7584 | 0,7584 | 0 | 0,6754 | 0,6754 | 1 |
| | Jornal 3 | 0,7638 | 0,7638 | 0 | 0,7535 | 0,7535 | 0 | 0,6451 | 0,6451 | 1 |
| Pureza | Reuters 1 | 0,5000 | 0,5000 | 0 | 0,5721 | 0,5721 | 0 | 0,8594 | 0,8594 | 1 |
| | Reuters 2 | 0,5136 | 0,5136 | 0 | 0,5457 | 0,5457 | 0 | 0,6538 | 0,6538 | 1 |
| | Reuters 3 | 0,3870 | 0,3870 | 0 | 0,4053 | 0,4053 | 0 | 0,5799 | 0,5799 | 1 |
| | Newsgroups 1 | 0,5205 | 0,5205 | 0 | 0,5670 | 0,5670 | 0 | 0,7220 | 0,7220 | 1 |
| | Newsgroups 2 | 0,3980 | 0,3980 | 0 | 0,4130 | 0,4130 | 0 | 0,5655 | 0,5655 | 1 |
| | Newsgroups 3 | 0,2933 | 0,2933 | 0 | 0,3250 | 0,3250 | 0 | 0,5683 | 0,5683 | 1 |
| | Jornal 1 | 0,4657 | 0,4657 | 0 | 0,5099 | 0,5099 | 1 | 0,5488 | 0,5488 | 1 |
| | Jornal 2 | 0,3977 | 0,3977 | 0 | 0,4220 | 0,4220 | 0 | 0,4724 | 0,4724 | 1 |
| | Jornal 3 | 0,3515 | 0,3515 | 0 | 0,3740 | 0,3740 | 0 | 0,4600 | 0,4600 | 1 |
| Rand | Reuters 1 | 0,5624 | 0,5624 | 0 | 0,6644 | 0,6644 | 0 | 0,8991 | 0,8991 | 1 |
| | Reuters 2 | 0,7999 | 0,7999 | 0 | 0,8154 | 0,8154 | 1 | 0,8543 | 0,8543 | 1 |
| | Reuters 3 | 0,8317 | 0,8317 | 1 | 0,7850 | 0,7850 | 0 | 0,8653 | 0,8653 | 1 |
| | Newsgroups 1 | 0,4987 | 0,4987 | 0 | 0,5170 | 0,5170 | 0 | 0,6227 | 0,6227 | 1 |
| | Newsgroups 2 | 0,5418 | 0,5418 | 0 | 0,5659 | 0,5659 | 0 | 0,6627 | 0,6627 | 1 |
| | Newsgroups 3 | 0,6139 | 0,6139 | 0 | 0,6445 | 0,6445 | 0 | 0,7758 | 0,7758 | 1 |
| | Jornal 1 | 0,5670 | 0,5670 | 0 | 0,6301 | 0,6301 | 1 | 0,6489 | 0,6489 | 1 |
| | Jornal 2 | 0,7263 | 0,7263 | 1 | 0,7509 | 0,7509 | 1 | 0,7758 | 0,7758 | 1 |
| | Jornal 3 | 0,8108 | 0,8108 | 1 | 0,8172 | 0,8172 | 1 | 0,8459 | 0,8459 | 1 |
| Silhueta | Reuters 1 | 0,0349 | 0,0914 | 0 | 0,0300 | 0,0000 | 0 | 0,0836 | 1,0000 | 1 |
| | Reuters 2 | 0,0426 | 0,0000 | 0 | 0,0561 | 0,3061 | 0 | 0,0867 | 1,0000 | 1 |
| | Reuters 3 | 0,0250 | 0,1013 | 0 | 0,0211 | 0,0000 | 0 | 0,0596 | 1,0000 | 1 |
| | Newsgroups 1 | 0,0194 | 0,3810 | 0 | 0,0170 | 0,0000 | 0 | 0,0233 | 1,0000 | 1 |
| | Newsgroups 2 | 0,0125 | 0,0323 | 0 | 0,0123 | 0,0000 | 0 | 0,0185 | 1,0000 | 1 |
| | Newsgroups 3 | 0,0102 | 0,0000 | 0 | 0,0120 | 0,1875 | 0 | 0,0198 | 1,0000 | 1 |
| | Jornal 1 | 0,0130 | 0,0000 | 0 | 0,0275 | 0,7474 | 0 | 0,0324 | 1,0000 | 1 |
| | Jornal 2 | 0,0186 | 0,0000 | 0 | 0,0263 | 0,4053 | 0 | 0,0376 | 1,0000 | 1 |
| | Jornal 3 | 0,0177 | 0,0000 | 0 | 0,0232 | 0,3459 | 0 | 0,0336 | 1,0000 | 1 |
| Davies-Bouldin | Reuters 1 | 4,3043 | 0,0000 | 0 | 1,7507 | 1,0000 | 1 | 1,9493 | 0,9922 | 0 |
| | Reuters 2 | 1,7877 | 1,0000 | 1 | 2,2385 | 0,0000 | 0 | 2,1084 | 0,2886 | 0 |
| | Reuters 3 | 3,1485 | 0,0000 | 0 | 1,8617 | 1,0000 | 1 | 2,3958 | 0,5849 | 0 |
| | Newsgroups 1 | 1,8431 | 1,0000 | 1 | 1,9543 | 0,8096 | 0 | 2,4270 | 0,0000 | 0 |
| | Newsgroups 2 | 1,9279 | 1,0000 | 1 | 1,9846 | 0,8096 | 0 | 2,8621 | 0,0000 | 0 |
| | Newsgroups 3 | 1,9372 | 1,0000 | 1 | 1,9916 | 0,9485 | 0 | 2,9943 | 0,0000 | 0 |
| | Jornal 1 | 1,8673 | 1,0000 | 1 | 1,8987 | 0,9712 | 1 | 2,9588 | 0,0000 | 0 |
| | Jornal 2 | 25,5900 | 0,0000 | 0 | 1,9444 | 1,0000 | 1 | 2,8176 | 0,9631 | 1 |
| | Jornal 3 | 26,1854 | 0,0000 | 0 | 2,8144 | 0,9995 | 1 | 2,8024 | 1,0000 | 1 |
| Dunn | Reuters 1 | 0,1162 | 0,0728 | 0 | 0,0879 | 0,0000 | 0 | 0,4768 | 1,0000 | 1 |
| | Reuters 2 | 0,0066 | 0,0000 | 0 | 0,0729 | 0,2296 | 0 | 0,2953 | 1,0000 | 1 |
| | Reuters 3 | 0,0102 | 0,0547 | 0 | 0,0000 | 0,0000 | 0 | 0,1866 | 1,0000 | 1 |
| | Newsgroups 1 | 0,3365 | 0,0000 | 0 | 0,3661 | 0,1326 | 0 | 0,5595 | 1,0000 | 1 |
| | Newsgroups 2 | 0,1890 | 0,0000 | 0 | 0,2462 | 0,2928 | 0 | 0,3843 | 1,0000 | 1 |
| | Newsgroups 3 | 0,1638 | 0,0000 | 0 | 0,2065 | 0,2568 | 0 | 0,3300 | 1,0000 | 1 |
| | Jornal 1 | 0,0817 | 0,0000 | 0 | 0,3729 | 0,7156 | 0 | 0,4887 | 1,0000 | 1 |
| | Jornal 2 | 0,0408 | 0,0000 | 0 | 0,1484 | 0,3247 | 0 | 0,3721 | 1,0000 | 1 |
| | Jornal 3 | 0,0367 | 0,0000 | 0 | 0,0978 | 0,2117 | 0 | 0,3255 | 1,0000 | 1 |
| **Total** | | | | **17** | | | **17** | | | **47** |

**Table 6** Parâmetros dos métodos iterativos

| Método | Descrição | Valor |
|---|---|---|
| *K-means* | Número de inicializações aleatórias de centróides | 10 |
| | Número máximo de iterações | 10 |
| Pesquisa harmônica | Número máximo de improvisações | 1.000 |
| | Número máximo de improvisações sem melhoria na função objetivo | 100 |
| | Percentual de consideração de soluções da memória harmônica | 0,45 |
| | Percentual mínimo de ajuste de acorde | 0,45 |
| | Percentual máximo de ajuste de acorde | 0,90 |
| | Fator determinante do tamanho da memória harmônica | 2 |
| Algoritmo genético | Probabilidade de recombinação do algoritmo genético | 0,8 |
| | Probabilidade de recombinação da evolução diferencial | 0,3 |
| | Probabilidade de mutação do algoritmo genético | 0,1 |
| | Probabilidade de mutação da evolução diferencial | 0,8 |
| | Número máximo de iterações sem melhoria na função objetivo | 100 |
| | Tamanho da população | 50 |

**Table 7** Análise comparativa dos métodos *K-means* e $C^3M$. RKM: Resultado *K-means*; EKM: Escore *K-means*; RC3M: Resultado $C^3M$; EC3M: Escore $C^3M$

| Critério | Conjunto de dados | RKM | EKM | RC3M | EC3M |
|---|---|---|---|---|---|
| Tempo de execução em segundos | NFS 1 | 51 | 0 | 0 | 1 |
| | NFS 2 | 891 | 0 | 5 | 1 |
| | NFS 3 | 3.108 | 0 | 17 | 1 |
| | Atos processuais 1 | 87 | 0 | 0 | 1 |
| | Atos processuais 2 | 1.618 | 0 | 9 | 1 |
| | Atos processuais 3 | 5.638 | 0 | 26 | 1 |
| Silhueta | NFS 1 | 0,2390 | 1 | 0,1625 | 0 |
| | NFS 2 | 0,1584 | 1 | 0,0966 | 0 |
| | NFS 3 | 0,1357 | 1 | 0,0604 | 0 |
| | Atos processuais 1 | 0,1988 | 1 | 0,0973 | 0 |
| | Atos processuais 2 | 0,1755 | 1 | 0,0110 | 0 |
| | Atos processuais 3 | 0,1735 | 1 | 0,0246 | 0 |
| Davies-Bouldin | NFS 1 | 1,3000 | 1 | 1,6497 | 0 |
| | NFS 2 | 1,6658 | 1 | 1,8531 | 0 |
| | NFS 3 | 1,8307 | 1 | 5,8719 | 0 |
| | Atos processuais 1 | 1,2758 | 1 | 1,9342 | 0 |
| | Atos processuais 2 | 1,2280 | 1 | 15,3854 | 0 |
| | Atos processuais 3 | 1,5317 | 1 | 70,0205 | 0 |
| Dunn | NFS 1 | 0,3217 | 1 | 0,0961 | 0 |
| | NFS 2 | 0,3386 | 1 | 0,0838 | 0 |
| | NFS 3 | 0,2438 | 1 | 0,0050 | 0 |
| | Atos processuais 1 | 0,3295 | 1 | 0,1192 | 0 |
| | Atos processuais 2 | 0,1300 | 1 | 0,0117 | 0 |
| | Atos processuais 3 | 0,0762 | 1 | 0,0017 | 0 |
| **Total** | | | **18** | | **6** |

**Table 8** Distribuição dos atos processuais com o emprego da representação *bag-of-words*

| Grupo | Atos processuais | Descritores |
|---|---|---|
| 1 | $a_1$, $a_2$, $a_6$, $a_8$, $a_{10}$, $a_{12}$, $a_{13}$, $a_{14}$, $a_{16}$, $a_{17}$, $a_{20}$, $a_{24}$, $a_{25}$, $a_{26}$, $a_{27}$, $a_{31}$, $a_{39}$, $a_{40}$, $a_{43}$, $a_{44}$, $a_{45}$, $a_{46}$, $a_{47}$, $a_{49}$, $a_{50}$, $a_{51}$, $a_{53}$, $a_{55}$, $a_{56}$, $a_{57}$, $a_{58}$, $a_{59}$, $a_{61}$, $a_{62}$, $a_{63}$, $a_{64}$, $a_{69}$, $a_{70}$, $a_{71}$, $a_{72}$, $a_{81}$, $a_{82}$, $a_{85}$, $a_{87}$, $a_{89}$, $a_{90}$, $a_{91}$, $a_{92}$, $a_{93}$, $a_{94}$, $a_{95}$, $a_{96}$, $a_{97}$, $a_{98}$ | cível, vara, oliveira, secretaria, danielle |
| 2 | $a_4$, $a_{18}$, $a_{19}$, $a_{21}$, $a_{22}$, $a_{23}$, $a_{28}$, $a_{29}$, $a_{30}$, $a_{32}$, $a_{33}$, $a_{34}$, $a_{35}$, $a_{36}$, $a_{37}$, $a_{38}$, $a_{41}$, $a_{42}$, $a_{54}$, $a_{60}$, $a_{66}$, $a_{67}$, $a_{68}$, $a_{73}$, $a_{74}$, $a_{83}$, $a_{86}$, $a_{88}$, $a_{100}$ | concretizado, anseia, afirmar, ressaltando-se, atestar |
| 3 | $a_{76}$, $a_{77}$, $a_{79}$, $a_{80}$ | itamar, notificações, spd, trt, costaoab |
| 4 | $a_5$, $a_7$, $a_9$, $a_{11}$, $a_{15}$, $a_{48}$, $a_{65}$, $a_{75}$, $a_{78}$, $a_{84}$, $a_{99}$ | digital, certificação, assinado, juíza, fevereiro |

**Table 9** Avaliação do algoritmo *K-means* aplicado à coleção de atos processuais com o emprego da representação *bag-of-words*

| Tempo de execução em milissegundos | Silhueta | Davies e Bouldin | Dunn |
|---|---|---|---|
| 35.473 | 0,1162 | 1,9399 | 0,6467 |

**Table 10** Conceitos utilizados na caracterização dos atos processuais

| Identificador | Descritores |
|---|---|
| Sentença | isso posto, face ao exposto, condeno, danos morais, danos materias, honorários advocatícios, honorários sucumbenciais |
| Audiência de Conciliação | audiência de conciliação |
| Audiência de Instrução e Julgamento | audiência de instrução e julgamento |
| Acórdão | nego provimento, dou provimento, provimento parcial, mantenho a sentença, provimento integral |
| Expedição de Alvará | expedição de alvará |
| Carta Precatória | carta precatória |

**Table 11** Distribuição dos atos processuais com o emprego da representação *bag-of-concepts*

| Grupo | Atos processuais | Descritores |
|---|---|---|
| 1 | $a_{18}$, $a_{23}$, $a_{24}$, $a_{25}$, $a_{26}$, $a_{27}$, $a_{28}$, $a_{29}$, $a_{30}$, $a_{31}$, $a_{32}$, $a_{33}$, $a_{34}$, $a_{35}$, $a_{36}$, $a_{37}$, $a_{38}$, $a_4$, $a_{41}$, $a_{42}$, $a_{43}$, $a_{46}$, $a_{48}$, $a_{49}$, $a_{50}$, $a_{51}$, $a_{55}$, $a_{56}$, $a_{57}$, $a_{58}$, $a_{59}$, $a_{62}$, $a_{63}$, $a_{73}$, $a_{74}$, $a_{76}$, $a_{77}$, $a_{79}$, $a_{80}$, $a_{82}$, $a_{83}$, $a_{88}$ | audiência de conciliação, acórdão, sentença |
| 2 | $a_{22}$, $a_{39}$, $a_{40}$, $a_{44}$, $a_{45}$, $a_{47}$, $a_{60}$, $a_{64}$, $a_{65}$, $a_{66}$, $a_{67}$, $a_{68}$, $a_{69}$, $a_{70}$ | carta precatória |
| 3 | $a_1$, $a_{53}$, $a_{61}$, $a_{75}$, $a_{78}$, $a_{81}$, $a_{85}$, $a_{87}$, $a_{89}$, $a_{92}$, $a_{93}$, $a_{95}$, $a_{96}$, $a_{97}$, $a_{98}$ | carta precatória, audiência de instrução e julgamento |
| 4 | $a_{100}$, $a_{54}$, $a_{71}$, $a_{72}$, $a_{84}$, $a_{86}$, $a_{90}$, $a_{91}$, $a_{94}$, $a_{99}$ | sentença, expedição de alvará |

**Table 12** Avaliação do algoritmo *K-means* aplicado à coleção de atos processuais com o emprego da representação *bag-of-concepts*

| Tempo de execução em milissegundos | Silhueta | Davies e Bouldin | Dunn |
|---|---|---|---|
| 100 | 0,6165 | 0,4576 | 0,2391 |

## 8. Conclusões

Em face da elevada e crescente quantidade de informações disponíveis em formato digital, a mineração de dados se apresenta como uma tecnologia inovadora, capaz de suplantar as dificuldades encontradas pelos métodos tradicionais de pesquisa e recuperação de informações, e de viabilizar a avaliação do conteúdo de grandes conjuntos de dados. Por meio do emprego de técnicas e algoritmos característicos, a mineração de dados suporta a extração de padrões originados de fontes de dados muitas vezes distribuídas e heterogêneas, proporcionando a obtenção de informações úteis e relevantes para diversas áreas do conhecimento. A mineração de textos, considerada uma especialização da tecnologia de mineração de dados, atua sobre extensas coleções de documentos aplicando processos de seleção, redução de dimensionalidade, representação matemática, mineração de dados, e verificação de resultados a fim de extrair do conteúdo dos textos informações relevantes para um determinado contexto de análise. Por atuar sobre fontes de dados desestruturadas e que demandam maior complexidade em seu processamento, as atividades de mineração de textos enfrentam dificuldades ainda mais desafiadoras do as que se apresentam às tarefas que se dedicam à mineração de dados em formato estruturado. A despeito dos grandes avanços realizados nas tecnologias de mineração de textos, observa-se que alguns aspectos das tarefas que compõem o seu processo ainda apresentam um certo grau de imprecisão, em decorrência de serem parcialmente dependentes do contexto do problema que está sendo abordado. Isto sugere que pode ser necessário, em algumas situações, aplicar mais de um conjunto de técnicas a fim de que resultados efetivamente relevantes e úteis sejam obtidos.

Neste sentido, este trabalho avaliou técnicas de mineração de textos, em particular aplicadas ao problema de agrupamento de atos processuais digitais em conjuntos de padrões semelhantes, no intuito de estabelecer um método apto em auxiliar a análise e interpretação das informações pertinentes aos textos publicados por meio dos diários de justiça eletrônicos. Uma análise dos métodos de agrupamento pesquisa harmônica, algoritmo genético, *K-means* e $C^3M$, sob a perspectiva de índices de validação de partições, estabeleceu o *K-means* como o algoritmo mais apropriado ao problema do particionamento de textos, e, tendo em conta que, no contexto deste estudo, o *K-means* era o único método que adotava como centróides a média dos elementos pertinentes aos grupos, sugere-se que este expediente seria mais apropriado do que a prática admitida pelos demais procedimentos, a qual consistia em determinar $K$ objetos da coleção como os centróides dos $K$ grupos. Considerando que o método *K-means* foi o melhor avaliado, a relevância que a escolha do número de grupos $K$ possui para este algoritmo e o desconhecimento da quantidade de grupos dentre os quais os atos processuais deveriam ser categorizados, o procedimento de particionamento de atos processuais sugerido incorporou o método de determinação de grupos do algoritmo IGN que, após operações de pré-processamento, estabelecia o valor de $K$ e o fornecia como entrada para o algoritmo *K-means*, que por sua vez determinava em quais grupos os textos deveriam ser dispostos. A operação de seleção dos descritores, que tinha como principal objetivo representar de maneira concisa o conteúdo dos grupos e por consequência facilitar a avaliação da coleção de documentos, não obteve resultados satisfatórios quando da utilização das palavras mais relevantes presentes nos atos processuais, haja vista que, embora significantes sob o aspecto estatístico, os termos selecionados eram desprovidos de importância semântica para o contexto de análise dos atos processuais. O modelo de representação *bag-of-concepts*, sugerido baseado em estudos anteriores e avaliado sob aspectos quantitativos e qualitativos, foi capaz de suplantar a deficiência verificada, e este resultado decorreu sobretudo do fato dos conceitos e seus identificadores serem definidos pelos próprios usuários, de modo orientado aos seus interesses. Ademais, a expressiva redução da quantidade de dimensões dos vetores utilizados para representar os textos, proporcionada pelo novo modelo, amplifica a capacidade dos algoritmos de particionamento em atuar sobre coleções de documentos mais extensas, sem que ocorra um incremento significativo da complexidade do processamento.

Como limitação do presente estudo refere-se o reduzido número de objetos presentes nas coleções de documentos textuais utilizadas nos experimentos de avaliação dos algoritmos de agrupamento, o qual foi decorrente da dificuldade em se obter, a partir da literatura, conjuntos de objetos previamente classificados e constituídos exclusivamente por textos.

## Author contributions

- Alfredo Silveira Araújo Neto : Concepção e desenho da pesquisa; Análise e interpretação dos resultados; Redação do manuscrito
- Marcos Negreiros Concepção: e desenho da pesquisa; Análise e interpretação dos resultados; Revisão crítica do manuscrito quanto ao conteúdo intelectual importante.

## Referências

[1] TAN, P. N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. 1. ed. Boston: Pearson Education, Inc., 2006. v. 1.

[2] GANTZ, J. F. et al. The diverse and exploding digital universe. *IDC – Anal. Future*, v. 1, n. 1, p. 1–14, 2008.

[3] REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Rev. Sist. Inf. FSMA*, v. 1, n. 7, p. 7–21, 2011.

[4] BALDAN, G. R. *Meio eletrônico: uma das formas de diminuição do tempo de duração do processo no 4º juizado especial de Porto Velho – RO*. Dissertação (Mestrado) — Fundação Getúlio Vargas, Rio de Janeiro, 2011.

[5] LEAL, A. C. de C. *A lei 11.419/2006 e a regulamentação das comunicações processuais eletrônicas no bojo do processo judicial telemático*. 2006. Online. Disponível em: ⟨http://jus.com.br/revista/texto/9298⟩.

[6] ORENGO, V.; HUYCK, C. A stemming algorithm for the portuguese language. In: *String Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings.Eighth International Symposium on*. Laguna de San Rafael, Chile: IEEE, 2001. '01.

[7] CAN, F.; OZKARAHAN, E. A. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. *ACM Trans. Database Syst. (TODS)*, v. 15, n. 4, p. 483–517, 1990.

[8] WANG, X.; QIU, W.; ZAMAR, R. H. Clues: A nonparametric clustering method based on local shrinking. *Comput. Stat. Data An.*, v. 52, n. 1, p. 286–298, 2007.

[9] VIANA, J. F. R.; GOMES, M. J. N.; XAVIER, A. F. S. Um algoritmo polinomial para identificação de grupos naturais em longas bases de dados. In: *XXXV Simpósio Brasileiro de Pesquisa Operacional*. Rio de Janeiro, Brasil: SOBRAPO, 2003. v. 1.

[10] MINER, G. et al. *Pratical Text Mining and Statistical Analysis for Non–structured text data applications*. 1. ed. Florida, USA: Elsevier, 2012.

[11] JAIN, A. K.; MURTY, M. N.; FLYNN, P. Data clustering: A review. *ACM Comput. Surv.*, v. 31, n. 3, p. 264–323, 1999.

[12] HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 1. ed. San Francisco: Morgan Kaufmann Publishers, 2006.

[13] JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. 1. ed. New Jersey: Prentice Hall, 1998.

[14] MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *An Introduction to Information Retrieval*. 1. ed. Cambridge: Cambridge University Press, 2009. v. 1.

[15] KALOGERATOS, A.; LIKAS, A. Text document clustering using global term context vectors. *Knowl. Inf. Syst.*, v. 31, n. 3, p. 455–474, 2012.

[16] KAROL, S.; MAGNAT, V. Evaluation of text document clustering approach based on particle swarm optimization. *Cent. Eur. J. Comput. Sci.*, v. 2, n. 3, p. 69–90, 2013.

[17] TSENG, Y.-H. Generic title labeling for clustered documents. *Expert Syst. Appl.*, v. 37, n. 3, p. 2247–2254, 2010.

[18] ZHANG, T. et al. Document clustering in correlation similarity measure space. *IEEE Trans. Knowl. Data Eng.*, v. 24, n. 6, p. 1002–1013, 2012.

[19] KALOGERATOS, A.; LIKAS, A. Document clustering using synthetic cluster prototypes. *Data Knowl. Eng.*, v. 70, n. 3, p. 284–306, 2011.

[20] LUO, C.; LI, Y.; CHUNG, S. M. Text document clustering based on neighbors. *Data Knowl. Eng.*, v. 68, n. 11, p. 1271–1288, 2009.

[21] ABUALIGAH, L. M.; KHADER, A. T.; HANANDEH, E. S. A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *J. Comput. Sci.*, v. 25, p. 456–466, 2018.

[22] KOBAYASHI, V. B. et al. Text mining in organizational research. *Organ. Res. Methods*, v. 21, n. 3, p. 733–765, 2017.

[23] JAIN, A. K. Data clustering: 50 years beyond k–means. *Pattern Recogn. Lett.*, v. 31, n. 8, p. 651–666, 2010.

[24] DRINEAS, P. et al. Clustering large graphs via the singular value decomposition. *Mach. Learn.*, v. 56, n. 3, p. 9–33, 2004.

[25] MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, California: University of California Press, 1967. v. 1.

[26] SLAMET, C. et al. Clustering the verses of the holy qur'an using k-means algorithm. *Asian J. Inf. Technol.*, v. 15, n. 24, p. 5159–5162, 2016.

[27] XIONG, C. et al. An improved k-means text clustering algorithm by optimizing initial cluster centers. In: *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. Macau, China: IEEE, 2016. (CCBD, v. 07).

[28] KANT, S.; ANSARI, I. A. An improved k–means clustering with atkinson index to classify liver patient dataset. *Int. J. Syst. Assur. Eng. Manag.*, v. 7, n. 1, p. 222–228, 2016.

[29] VINUÉ, G.; SIMÓ, A.; ALEMANY, S. The k–means algorithm for 3d shapes with an application to apparel design. *Adv. Data Anal. Classi.*, v. 10, n. 1, p. 103–132, 2016.

[30] GANESH, M.; NARESH, M.; ARVIND, C. Mri brain image segmentation using enhanced adaptive fuzzy k-means algorithm. *Intell. Autom. Soft Comput.*, v. 23, n. 2, p. 325–330, 2017.

[31] BAI, L. et al. Fast density clustering strategies based on the k–means algorithm. *Pattern Recogn.*, v. 71, n. Supplement C, p. 375–386, 2017.

[32] FORSATI, R. et al. Efficient stochastic algorithms for document clustering. *Inform. Sciences*, v. 220, n. 1, p. 269–291, 2013.

[33] ALIA, O.; MANDAVA, R. The variants of the harmony search algorithm: an overview. *Artif. Intell. Rev.*, v. 36, n. 1, p. 49–68, 2011.

[34] GEEM, Z. W. Particle–swarm harmony search for water network design. *Eng. Optimiz.*, v. 41, n. 4, p. 297–311, 2009.

[35] VERMA, A.; PANIGRAHI, B.; BIJWE, P. Harmony search algorithm for transmission network expansion planning. *IET Gener. Transm. Distrib.*, v. 4, n. 6, p. 663–673, 2010.

[36] RAZFAR, M. R.; ZINATI, R. F.; HAGHSHENAS, M. Optimum surface roughness prediction in face milling by using neural network and harmony search algorithm. *Int. J. Adv. Manuf. Technol.*, v. 52, n. 5, p. 487–495, 2011.

[37] LE, D. L.; HO, D. L.; VO, N. D. Hybrid differential evolution and harmony search for optimal power flow. *Glob. J. Technol. Optim.*, v. 6, n. 2, p. 1–15, 2015.

[38] AFSHAR, M. H. et al. Exploring the efficiency of harmony search algorithm for hydropower operation of multi-reservoir systems: A hybrid cellular automat-harmony search approach. In: SER, J. D. (Ed.). *Harmony Search Algorithm*. Singapore: Springer Singapore, 2017. (AISC, v. 514).

[39] NIGDELI, S. M.; BEKDAŞ, G.; YANG, X.-S. Optimum tuning of mass dampers by using a hybrid method using harmony search and flower pollination algorithm. In: SER, J. D. (Ed.). *Harmony Search Algorithm*. Singapore: Springer Singapore, 2017. (AISC, v. 514).

[40] MAHDAVI, M.; ABOLHASSANI, H. Harmony k–means algorithm for document clustering. *Data Min. Knowl. Disc.*, v. 18, n. 3, p. 370–391, 2009.

[41] ALIA, O. M.; MANDAVA, R.; AZIZ, M. E. A hybrid harmony search algorithm for mri brain segmentation. *Evol. Intell.*, v. 4, n. 1, p. 31–49, 2011.

[42] MEENA, Y. K.; SHASHANK; SINGH, V. P. Text documents clustering using genetic algorithm and discrete differential evolution. *Int. J. Comput. Appl.*, v. 43, n. 1, p. 16–19, 2012.

[43] LINDEN, R. *Algoritmos Genéticos*. 1. ed. Rio de Janeiro: Brasport, 2008. v. 1.

[44] MELANIE, M. *An Introduction to Genetic Algorithms*. 1. ed. Cambridge, Massachusetts: MIT Press, 1996.

[45] VIANA, V. *Meta–heurísticas e Programação Paralela em Otimização Combinatória*. 1. ed. Fortaleza: EUFC, 1998. v. 1.

[46] THEDE, S. M. An introduction to genetic algorithms. *J. Comput. Sci. Coll.*, v. 20, n. 1, p. 115–123, 2004.

[47] RAMPAZZO, P. C. B. *Planejamento hidrelétrico: otimização multiobjetivo e abordagens evolutivas*. Tese (Doutorado) — Universidade Estadual de Campinas, Campinas, 2012.

[48] PORTER, M. F. Readings in information retrieval. In: JONES, K. S.; WILLETT, P. (Ed.). 1. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. v. 1, cap. An Algorithm for Suffix Stripping, p. 313–316.

[49] CONRAD, J. G. et al. Effective document clustering for large heterogeneous law firm collections. In: SARTOR, G. (Ed.). *Proceedings of the 10th International Conference on Artificial Intelligence and Law*. Bologna, Italy: ACM, 2005. (ICAIL, '05).

[50] STEIN, B.; EISSEN, S. M. zu; WIBBROCK, F. On cluster validity and the information need of users. In: *3rd Int. Conference on Artificial Intelligence and Applications*. Calgary, AB, Canada: ACTA Press, 2003. v. 1.

[51] DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *J. Cybernetics*, v. 3, n. 3, p. 32–57, 1973.

[52] WIVES, L. K. *Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos*. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

[53] SHEHATA, S.; KARRAY, F.; KAMEL, M. Enhancing text clustering using concept-based mining model. In: *Data Mining, 2006. ICDM 06. Sixth International Conference on*. New York, NY, USA: IEEE, 2006. (ICDM, '06).

[54] BAGHEL, R.; DHIR, R. A frequent concepts based document clustering algorithm. *Int. J. Comput. Appl.*, v. 4, n. 5, p. 6–12, 2010.

[55] PAPADIMITRIOU, C. H. et al. Latent semantic indexing: A probabilistic analysis. *J. Comput. Syst. Sci.*, v. 61, n. 2, p. 217–235, 2000.

[56] MEI, Q.; SHEN, X.; ZHAI, C. Automatic labeling of multinomial topic models. In: BERKHIN, P.; CARUANA, R.; WU, X. (Ed.). *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2007. (KDD, '07).

[57] HONG, L.; DAVISON, B. D. Empirical study of topic modeling in twitter. In: MELVILLE, P.; LESKOVEC, J.; PROVOST, F. (Ed.). *Proceedings of the First Workshop on Social Media Analytics*. New York, NY, USA: ACM, 2010. (SOMA, '10).

[58] GAO, W.; LI, P.; DARWISH, K. Joint topic modeling for event summarization across news and social media streams. In: CHEN, X. et al. (Ed.). *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2012. (CIKM, '12).

[59] SAHLGREN, M.; COSTER, R. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In: *Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2004. (COLING, '04).

[60] ALAHMADI, A.; JOORABCHI, A.; MAHDI, A. E. A new text representation scheme combining bag-of-words and bag-of-concepts approaches for automatic text classification. In: *2013 7th IEEE GCC Conference and Exhibition*. Doha, Qatar: IEEE, 2013. (GCC, '13).

[61] GARCÍA, M. A. M.; RODRÍGUEZ, R. P.; RIFÓN, L. E. A. Biomedical literature classification using encyclopedic knowledge: a wikipedia-based bag-of-concepts approach. *PeerJ*, v. 3, n. 1, p. e1279, 2015.