Avaliação da performance de índices de similaridade aplicados ao agrupamento de objetos textuais

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

Resumo: A captura e o 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. Neste contexto, as atividades de análise de agrupamentos ou classificação não supervisionada de objetos, se constituem como uma das técnicas de mineração de informações mais frequentemente empregadas no intuito de proporcionar a organização do volume progressivamente crescente de elementos textuais, por meio da disposição dos documentos em grupos de itens semelhantes com base em um índice de similaridade. Neste sentido, este estudo avalia os índices de similaridade distância Euclidiana, distância do coseno, distância de Hamming, coeficiente de Jaccard estendido e coeficiente de correlação de Pearson, sob a perspectiva de seis índices de validação de agrupamentos, observando que a distância do coseno representa, conforme a presente análise, o índice de similaridade mais apropriado ao agrupamento de objetos textuais, convertidos em formato estruturado por intermédio de técnicas de mineração de textos.

***Palavras-chave*:** Análise de agrupamentos. Agrupamento de documentos. Índices de similaridade.

Abstract: *The capture and the digital data store have allowed companies the accumulation of an extremely high volume of information, constituted mainly by unstructured data, represented by texts. In this context, the cluster analysis operations or unsupervised classification of objects, represent one of the most frequently used data mining techniques to provide the organization of the progressively increasing volume of textual elements, by means of arrangement of the documents in groups of similar itens based in a similarity measure . In this sense, this article evaluate the similarity measures Euclidiean distance, cosine distance, Hamming distance, extended Jaccard coefficient and Pearson's correlation coefficient, from the perspective of six clustering validation indexes, noticing that the cosine distance represent, according to this analysis, the similarity measure most appropriate to clustering textual objects, converted into structured format through text mining techniques.*

***Keywords*:** *Clustering analysis. Document clustering. Similarity index*

1 Introdução

A mineração de dados é um processo de descoberta automática de conhecimento em grandes repositórios de dados. Correspondente a um conjunto de técnicas que atuam sobre grandes bancos de dados a fim de identificar padrões úteis que, de outra forma, permaneceriam desconhecidos. As tarefas da mineração de dados são classificadas em duas categorias principais: tarefas de previsão e tarefas descritivas. As tarefas de previsão têm como objetivo prever o conteúdo de um determinado atributo, nomeado como a variável dependente ou alvo, com base nos valores de outros atributos, conhecidos como variáveis independentes ou explicativas. Já as tarefas descritivas

[1]Techway Informática Ltda., Av. Dom Luis, 500 - Sala 1709, Fortaleza - Ceará, Brasil
{alfredosilveira@yahoo.com.br}
[2]Universidade Estadual do Ceará, Mestrado Profissional em Computação, Av. Paranjana, 1700 - Campus do Itaperi, Fortaleza - Ceará, Brasil
{negreiro@graphvs.com.br}

http://dx.doi.org/10.5335/v9i4.7082

têm como propósito derivar padrões (correlações, tendências, grupos, trajetórias e anomalias) que sintetizem relacionamentos subjacentes nos dados. São tarefas de natureza exploratória que não raramente demandam a aplicação de técnicas de pós-processamento a fim de validar e esclarecer os resultados obtidos [1].

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. 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. A análise de agrupamentos tem sido amplamente utilizada em muitas aplicações, a exemplo da pesquisa de mercado, do reconhecimento de padrões, da análise de dados ou do processamento de imagens. Na área de negócios, a análise de agrupamentos é capaz de auxiliar os empresários a descobrir grupos distintos em suas bases de clientes, que podem ser caracterizados conforme os padrões de compra que apresentam. 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 classificação de documentos oriundos da *web*, no intuito de promover a descoberta de informações relevantes implícitas nos conteúdos dos textos [1, 2, 3].

De acordo com [4], a maior parte dos textos disponíveis para acesso por intermédio de livros, jornais, mensagens eletrônicas ou páginas *web* é constituída por elementos não estruturados ou semi-estruturados. Os itens não estruturados são expressos por textos que não foram submetidos a nenhum tipo de processamento, e que por consequência apresentam-se em seu formato original, ao passo que os textos semi-estruturados são, por exemplo, representados pelo conteúdo de uma planilha eletrônica, ou ainda por um atributo de uma tabela de banco de dados. Ainda segundo [4], estimativas indicam que entre 75 e 80% dos dados mundialmente disponíveis são compostos por textos em formato não estruturado ou semi-estruturado, circunstância esta que pode ser observada considerando-se apenas o volume de informações não estruturadas, armazenadas por intermédio de páginas eletrônicas da Internet. Com efeito, entre 2001 e 2009, o número de páginas *web* disponíveis por meio da rede mundial de computadores acresceu de aproximadamente 10 milhões para cerca de 150 bilhões. Desta forma, embora o índice de expansão verificado não seja linear e não necessariamente a mesma taxa de crescimento verificada seja ainda conservada, a quantidade de informações registradas por meio de páginas eletrônicas denota o quão rápido dados textuais têm sido internacionalmente acumulados.

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 não estruturado 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 volume de dados textuais exponencialmente crescente, a fim de obter dos mesmos informações relevantes [4].

As dificuldades decorrentes do crescente acúmulo de documentos eletrônicos têm sido abordadas por intermédio de numerosas técnicas de mineração de textos, e, dentre elas, destacam-se àquelas relacionadas às atividades de agrupamento de documentos. Os algoritmos de agrupamento de documentos, que têm como propósito dispor os arquivos em grupos de forma que os textos que façam parte de um dado grupo sejam mais semelhantes entre si do que os textos que pertencem a grupos distintos, possibilitam que extensas coleções de objetos textuais sejam eficientemente organizadas e sumarizadas, e contribuem para que as consultas realizadas sobre os elementos destas coleções possam ser executadas de forma mais célere e eficaz. Em geral, uma técnica de agrupamento apropriada tem como objetivo proporcionar a minimização da distância entre os elementos que fazem parte de um grupo (distância intra-grupo), ao mesmo tempo em que procura maximizar a distância entre os objetos que pertencem a diferentes grupos (distância inter-grupo), desta forma, verifica-se que a medida ou o índice de similaridade empregado na operação de determinação do afastamento entre os objetos, exerce significante influência nos resultados alcançados. Variados métodos capazes de estabelecer a semelhança entre objetos têm sido propostos e aplicados na literatura, muito embora a eficácia destes modelos, quanto empregados por algoritmos de agrupamento de documentos, ainda seja discutível [5, 6, 7].

A despeito da existência de diversos estudos [8, 9, 10, 11, 12, 13, 14, 15, 16] que avaliaram a eficácia de índices de similaridade aplicados à operação de agrupamento de objetos textuais, a presente análise estende estes trabalhos mediante o exame empírico de cinco índices de semelhança distintos, com o emprego de seis índices de validação de resultados. Em particular, os índices de similaridade distância Euclidiana, distância do coseno,

distância de Hamming, coeficiente de Jaccard estendido e coeficiente de correlação de Pearson, foram utilizados para realizar o agrupamento de nove conjuntos de documentos de diferentes extensões e características, com a aplicação do método de particionamento *k-means*. As soluções obtidas foram subsequentemente avaliadas por intermédio dos índices de validação de agrupamento Entropia, Pureza, Rand, Ajusted Rand, Fowlkes-Mallows e Γ Statistic, no intuito de que o índice de semelhança mais apropriado pudesse ser indicado.

Além da seção 1, representada pela presente introdução, este trabalho compreende cinco seções adicionais, as quais estão organizadas conforme especificado a seguir. A seção 2 descreve a atividade de análise de agrupamentos, apresenta índices de proximidade que podem ser utilizados na determinação do agrupamento de objetos e caracteriza os índices de validação que podem ser empregados na avaliação dos métodos de particionamento. As seções 3 e 4 descrevem, respectivamente, o problema do agrupamento de objetos textuais e um dos métodos de particionamento que pode ser aplicado na sua resolução. A seção 5 especifica os experimentos de avaliação que foram conduzidos com o propósito de comparar os índices de similaridade, e, por fim, a seção 6 refere as conclusões obtidas conforme os resultados dos experimentos.

2 Análise de agrupamentos

Uma das mais importantes operações relacionadas à análise de dados, consiste em classificar ou agrupar os objetos em um conjunto de categorias ou grupos, de forma que os elementos relacionados a um mesmo grupo apresentem características equivalentes, de acordo com um determinado critério. Com efeito, verifica-se que a classificação desempenha um importante e indispensável papel no desenvolvimento humano, haja vista que para melhor compreender um novo objeto ou fenômeno, as pessoas frequentemente procuram identificar características descritivas dos mesmos, no intuito de comparará-las com aquelas que pertencem a objetos ou fenômenos conhecidos [17].

Fundamentalmente, os sistemas de classificação podem ser decompostos em sistemas supervisionados e sistemas não supervisionados. A classificação supervisionada emprega uma coleção de objetos previamente qualificados, e o problema consiste em agrupar novos objetos, ainda não rotulados, com base nas informações obtidas por meio dos elementos já classificados. Na classificação não supervisionada, também denominada de análise de agrupamentos ou análise exploratória de dados, não existem rótulos de dados dispóniveis, e o objetivo é decompor uma coleção finita de objetos não categorizados, em um conjunto finito de grupos "naturais". O emprego de técnicas de análise de agrupamentos provém da necessidade de se investigar dados de natureza desconhecida sem que exista qualquer conhecimento prévio acerca dos mesmos, promovendo a divisão dos objetos em uma coleção de subgrupos relativamente homogêna, com base em uma medida de similaridade, ocasionalmente subjetiva. Diferentes critérios de agrupamento, algoritmos de agrupamento distintos, ou ainda parâmetros distintos empregados para o mesmo algoritmo, podem resultar em partições completamente desiguais, a exemplo dos seres humanos, que quando classificados conforme sua etnia, região, situação econômica, grau de instrução, peso ou altura, podem originar grupos integralmente distintos [17, 18].

2.1 Índices de similaridade

Um aspecto de fundamental importância na tentativa de identificar grupos de padrões que possam estar presentes nos conjuntos de dados, consiste na determinação do quão próximo os objetos apresentam-se uns dos outros ou o quão distante eles estão. Muitos dos métodos de análise de agrupamentos têm como fundamento uma matriz $n \times n$, onde n representa o número de objetos da coleção de padrões, cujos elementos refletem uma medida quantitativa de proximidade, mais frequentemente referenciada como dissimilaridade, distância ou similaridade. Dois objetos são considerados próximos quando sua dissimilaridade ou distância é reduzida e sua similaridade é elevada. Dado um par constituído por dois padrões i e k quaisquer, pertencentes a uma determinada coleção de objetos O, a similaridade ou semelhança entre os elementos, denotada por $d(i, k)$, pode ser calculada mediante um índice de proximidade que satisfaz às seguintes propriedades: (i) $d(i, i) = 0, \forall\, i \in O$, para uma dissimilaridade, ou $d(i, i) \geq max\ d(i, k), \forall\, i, k \in O$, para uma similaridade; (ii) $d(i, k) = d(k, i), \forall\, i, k \in O$; (iii) $d(i, k) \geq 0, \forall\, i, k \in O$ [19, 18].

2.1.1 Distância de Minkowski

Um índice de proximidade pode ser computado de diversas maneiras. Uma medida de dissimilaridade habitualmente utilizada para calcular a diferença entre os elementos de uma matriz de padrões $[x_{ij}]$ de atributos contínuos, na qual x_{ij} representa a j-ésima característica do i-ésimo padrão, consiste na distância de Minkowski. Se o i-ésimo padrão, o qual representa a i-ésima linha da matriz de padrões, for denotado pelo vetor $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T$, $i = 1, 2, \ldots, n$, com m significando o número de características, n o número de padrões, e T o vetor transposto, então a distância de Minkowski será definida por $d(i,k) = \left[\sum_{j=1}^{m} \mid x_{ij} - x_{kj} \mid^r\right]^{(1/r)}$, $\forall\, i, k \in O$ onde $r \geq 1$. Para todas as distâncias de Minkowski as propriedades adicionais estabelecidas a seguir, as quais determinam que uma dissimilaridade representa uma métrica, são satisfeitas: (i) $d(i,k) = 0 \Leftrightarrow x_i = x_k$, $\forall\, i, k \in O$; (ii) $d(i,k) \leq d(i,l) + d(l,k)$, $\forall\, i, k, l \in O$ (desigualdade triangular) [18].

De acordo com [1, 18, 20], os três exemplos mais comuns de métricas de Minkowski são: (i) Manhattan, para $r = 1$, $d(i,k) = \sum_{j=1}^{m} \mid x_{ij} - x_{kj} \mid$, $\forall\, i, k \in O$; (ii) Distância Euclidiana, para $r = 2$, $d(i,k) = \left[\sum_{j=1}^{m}(x_{ij} - x_{kj})^2\right]^{(1/2)} = \left[(x_i - x_k)^T(x_i - x_k)\right]^{1/2}$, $\forall\, i, k \in O$; (iii) Distância suprema, para $r \to \infty$, $d(i,k) = \lim_{r\to\infty}\left[\sum_{j=1}^{m}(x_{ij} - x_{kj})^r\right]^{(1/r)}$, $\forall\, i, k \in O$. O parâmetro r não deve ser confundido com o número de atributos d dos vetores que caracterizam os objetos. As distâncias Manhatthan, Euclidiana e suprema estão definidas para todos os valores de $m = 1, 2, 3, \ldots$, e determinam formas distintas de associar cada dimensão dos padrões em uma medida de distância geral. Quando os padrões possuem o mesmo número de características e são representados por atributos de conteúdo binário, a métrica de Manhattan é denominada distância de Hamming, representa o número de dimensões nos quais os padrões que estão sendo comparados diferem, e é determinada por intermédio da expressão $d(i,k) = \sum_{j=1}^{m} H(x_{ij}, x_{kj})$, $\forall\, i, k \in O$ onde $H(x_{ij}, x_{kj}) = 0$ se $x_{ij} = x_{kj}$, e $H(x_{ij}, x_{kj}) = 1$ quando $x_{ij} \neq x_{kj}$.

2.1.2 Distância do coseno

No contexto da análise de agrupamentos, documentos são habitualmente representados como vetores nos quais cada atributo pode significar, por exemplo, o número de vezes em que um termo (palavra) em particular ocorre no documento. Embora os documentos de uma coleção possam encerrar milhares ou dezenas de milhares de atributos, cada elemento é descrito por um vetor esparso, pois a quantidade de atributos com valores diferentes de zero é relativamente reduzida. Desta forma, a medida de similaridade entre estes tipos de padrões não deve ser dependente do número de características com valor zero compartilhadas, desde que quaisquer dois documentos provavelmente não compreenderão muitas palavras em comum, e, por consequência, se as correspondências 0-0 forem consideradas a maioria dos documentos será demasiadamente similar à maioria dos outros documentos [1].

Tendo em vista que uma medida de semelhança entre documentos deve desconsiderar correspondências 0-0, a similaridade do coseno definida a seguir, é uma das medidas de semelhança de documentos mais regularmente utilizadas. Se i e k representam dois vetores de documentos, então $cos(i,k) = \frac{i \cdot k}{\|i\| \cdot \|k\|}$, $\forall\, i, k \in O$ onde $\cdot$ indica o produto escalar, $i \cdot k = \sum_{j=1}^{m} i_j k_j$ e $\| i \|$ é o comprimento do vetor i, $\| i \| = \sqrt{\sum_{j=1}^{m} i_j^2} = \sqrt{i \cdot i}$. A similaridade do coseno entre dois documentos i e k consiste na verdade em uma medida do coseno do ângulo entre os vetores que os representam. Desta forma, a distância ou dissimilaridade do coseno entre os dois vetores normalizados será computada por $d(i,k) = 1 - cos(i,k) = 1 - \frac{i \cdot k}{\|i\| \cdot \|k\|}$, $\forall\, i, k \in O$. [1, 21].

De fato, se a similaridade do coseno for 1, então a dissimilaridade será 0 e o ângulo entre i e k será igual a 0°, com i e k correspondendo ao mesmo documento. De modo contrário, se a similaridade do coseno for 0, então dissimilaridade será 1 e o ângulo entre os vetores será 90°, não existindo portanto qualquer termo em comum entre os documentos [1].

2.1.3 Coeficiente de Jaccard estendido

O coeficiente de Jaccard estendido, o qual pode ser utilizado para estabelecer a similaridade entre objetos textuais e que é também conhecido como coeficiente de Tanimoto, é definido pela expressão $J(i,k) = \frac{i \cdot k}{\|i\|^2 + \|k\|^2 - i \cdot k}$, $\forall\, i, k \in O$ onde $\cdot$ indica o produto escalar, $i \cdot k = \sum_{j=1}^{m} i_j k_j$ e $\| i \|$ é o comprimento do vetor i,

$\| i \| = \sqrt{\sum_{j=1}^{m} i_j^2} = \sqrt{i \cdot i}$. Tendo em conta que, de modo inverso ao que ocorre com a distância Euclidiana, o coeficiente de Jaccard estendido representa uma similaridade limitada ao intervalo $[0, 1]$, pratica-se a transformação representada pela expressão $d(i, k) = 1 - J(i, k)$, no intuito de que o valor do coeficiente entre dois objetos i e k quaisquer, corresponda a uma distância [1, 6].

2.1.4 Coeficiente de correlação de Pearson

De acordo com [1], o coeficiente de correlação, que denota o relacionamento linear entre objetos constituídos por atributos binários ou contínuos, é determinado pela equação $c(i, k) = \frac{s_{ik}}{s_i s_k}$, $\forall\ i, k \in O$ onde $s_{ik} = \frac{1}{n-1} \sum_{j=1}^{m} (i_j - m^{(i)})(k_j - m^{(k)})$, $s_i = \sqrt{\frac{1}{n-1} \sum_{j=1}^{m} (i_j - m^{(i)})^2}$, $s_k = \sqrt{\frac{1}{n-1} \sum_{j=1}^{m} (k_j - m^{(k)})^2}$, $m^{(i)} = \frac{1}{n} \sum_{j=1}^{m} i_j$ e $m^{(k)} = \frac{1}{n} \sum_{j=1}^{m} k_j$.

O valor do coeficiente de correlação situa-se no intervalo $[-1, 1]$, indicando que há entre i e k um perfeito relacionamento linear negativo ou positivo, conforme o resultado auferido seja, nesta ordem, -1 ou 1. Valores iguais a 0 denotam que não existe qualquer relacionamento linear entre os atributos dos objetos, embora um relacionamento não-linear possa ser eventualmente verificado. Desde que o resultado do coeficiente de correlação caracteriza uma similaridade, admite-se a transformação designada pela expressão $d(i, k) = 1 - c(i, k)$, quando $c(i, k) \geq 0$, ou $d(i, k) = |c(i, k)|$, para $c(i, k) < 0$, a fim de que o valor obtido represente uma dissimilaridade [1, 22].

2.2 Índices de validação de agrupamentos

A avaliação do resultado de um agrupamento, em geral realizada por meio da utilização de critérios estatísticos capazes de expressar a qualidade das estruturas obtidas, é certamente uma atividade relevante haja vista que a aplicação dos métodos de agrupamento sobre um conjunto de dados sempre resultará na divisão dos padrões entre os grupos, mesmo que estes sejam irreais ou ainda de pouca significância para o contexto do problema que se deseja determinar. A validade de uma estrutura de grupos pode ser expressa em termos de três tipos de critérios: externos, internos e relativos. Os critérios externos analisam a performance comparando a estrutura de grupos alcançada com uma informação previamente conhecida. Os critérios internos avaliam a adequação entre a estrutura de grupos e os dados, utilizando somente informações pertinentes aos próprios dados. Por fim, os critérios relativos comparam estruturas de agrupamento entre si a fim de determinar quais são mais estáveis ou mais apropriadas aos dados [2, 17].

2.2.1 Entropia e pureza

De acordo com [23], a entropia e a pureza representam critérios de avaliação externos que usam o conhecimento prévio das classes às quais os elementos do conjuntos de dados estão associados. O cálculo da entropia é dependente da distribuição das classes dentre os constituintes do agrupamento resultante, enquanto que a pureza é sujeita ao tamanho relativo da maior classe concernente ao resultado do agrupamento. Dado um grupo particular c_k, de tamanho n_k, a entropia do k-ésimo grupo é definida como $e_k = -\frac{1}{\log_2 c} \sum_{i=1}^{c} \frac{n_k^i}{n_k} \log_2 \frac{n_k^i}{n_k}$, onde c é o número de classes do conjunto de dados e n_k^i e número de elementos de i-ésima classe que foram associados ao k-ésimo grupo. A entropia total da partição é então definida como a soma das entropias de cada grupo ponderada pelo número de elementos do mesmo, ou seja $E_k = \sum_{k=1}^{K} \frac{n_k}{n}\ e_k$. Em geral, quanto menor a entropia maior será a qualidade do agrupamento, de sorte que uma partição de grupos perfeita será aquela que resultar em grupos constituídos por elementos de uma única classe, situação na qual obtém-se valor de entropia igual a zero. De modo similar, a pureza de um grupo c_k de tamanho n_k é estabelecida como $p_k = \frac{1}{n_k} max_i(n_k^i)$, na qual o número de elementos da classe mais frequente no grupo é dividido pelo número de elementos do grupo. A pureza de uma partição, obtida tomando-se a soma ponderada da pureza de cada grupo em particular, é determinada como $P_k = \sum_{k=1}^{K} \frac{n_k}{n} p_k$.

Agrupamentos de qualidade inferior possuem valores de pureza próximos a zero, ao passo que agrupamentos aproximadamente perfeitos possuem valores de pureza próximos a 1. Valores elevados de pureza podem ser alcançados se o número de grupos for alto. Em particular, obtém-se valor de pureza igual 1 se cada padrão for atribuído ao seu próprio grupo, e, em razão desta característica, não se recomenda o emprego da pureza na avaliação

da qualidade dos agrupamentos em função do número de grupos selecionado pelo método de agrupamento [24].

2.2.2 Rand, Fowlkes-Mallows, Γ Statistic e Ajusted Rand

Os índices Rand, Fowlkes-Mallows, Γ Statistic e Ajusted Rand constituem critérios de avaliação externos determinados a partir de uma tabela de contingência construída com base na comparação entre os números dos grupos atribuídos aos padrões pelo algoritmo de agrupamento e os rótulos de categoria anteriormente associados aos dados [18].

Se $U = \{u_i\}, i = 1, \ldots, R$ e $V = \{v_j\}, j = 1, \ldots, C$ representarem duas partições de n objetos, onde U corresponde à solução resultante da aplicação do método de agrupamento e V consiste no agrupamento previamente definido sobre o conjunto de dados, então a entrada n_{ij} da tabela 1 será o número de objetos presentes simultâneamente nos grupos u_i e v_j. O termo $n_{i.}$ será a soma dos elementos da i-ésima linha ou número de objetos do grupo u_i, e $n_{.j}$ indicará a soma dos elementos da j-ésima coluna ou número de objetos que pertencem ao grupo v_j [18].

Tabela 1: Contingência entre duas partições

	v_1	v_2	v_3	v_4	$\ldots$	v_C	
u_1	n_{11}	n_{12}	n_{13}	n_{14}	$\ldots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	n_{23}	n_{24}	$\ldots$	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_r	n_{R1}	n_{R2}	n_{R3}	n_{R4}	$\ldots$	n_{RC}	$n_{R.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$\ldots$	$n_{.C}$	$n_{..} = n$

Os índices de avaliação externos que podem ser determinados em termos dos valores apresentados na tabela 1, também podem ser expressos por meio das seguintes funções de indicação, demonstrando a similaridade entre estes índices e as medidas de proximidade para objetos constituídos de atributos binários: (i) $I_U(i, j) = 1$ se $x_i \in u_r$ e $x_j \in u_r$, para $1 \leq r \leq R$ ou $I_U(i, j) = 0$, caso contrário; (ii) $I_V(i, j) = 1$ se $x_i \in v_s$ e $x_j \in v_s$, para $1 \leq s \leq C$ ou $I_V(i, j) = 0$, caso contrário. Se U representar a partição obtida pelo algoritmo de agrupamento e V corresponder ao agrupamento previamente definido sobre o conjunto de dados, então $I_U(i, j)$ terá valor 1 quando os objetos x_i e x_j estiverem no mesmo grupo, e 0 quando os objetos pertencerem a grupos distintos. Analogamente, $I_V(i, j)$ será 1 se os dois objetos estiverem na mesma categoria e 0, caso contrário. O diagrama de contingência para as funções de indicação, representado pela tabela 2 demonstra o número de maneiras por meio das quais os pares de objetos podem ser analisados pelas duas partições [18].

Tabela 2: Contingência para as funções de indicação

		I_V		
		1	0	
I_U	1	a	b	m_1
	0	c	d	$M - m_1$
		m_2	$M - m_2$	

Por conseguinte, a corresponderá ao número de pares de objetos que estão nos mesmos grupos nas duas partições, d será o número de pares que estão em diferentes grupos em ambas as partições, e b o número de pares que estão no mesmo grupo em U, mas em diferentes grupos na partição V. O número de pares de objetos que estão no mesmo grupo em U será denotado por $m_1 = a + b$, ao passo que $m_2 = a + c$ corresponderá ao número de pares de objetos situados no mesmo grupo em V. O número total de pares de objetos considerando U e V será expresso por $M = a + b + c + d = \frac{n(n-1)}{2}$. Com base nas tabelas de contingência, os índices de validação externos Rand (R), Fowlkes-Mallows (FM), e Γ Statistic (Γ) são determinados, respectivamente, pelas expressões $R = \frac{(a+d)}{\binom{n}{2}}$,

$FM = \frac{a}{\sqrt{m_1 m_2}}$, e $\Gamma = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 (M-m_1)(M-m_2)}}$ [18].

Valores elevados para os índices Fowlkes-Mallows e Rand, implicam em considerável semelhança entre as duas partições. De modo análogo ao coeficiente de correlação de Pearson, o índice Γ Statistic consiste em uma correlação, apresentando resultados compreendidos entre -1 e 1, onde valores negativos sugerem uma discordância entre as partições, enquanto que valores positivos e próximos de 1 indicam uma perfeita concordância. Os índices Fowlkes-Mallows e Rand possuem valores situados entre 0 e 1, com 1 denotando que as partições são indênticas e 0 indicando que nenhum par de objetos ocorre no mesmo grupo ou em grupos distintos, em ambas as partições [18, 25, 26].

Valores iguais a 0 para o índice Rand ocorrem somente em situações nas quais uma partição consiste de um único grupo que abrange todos os objetos, enquanto que valores iguais a 1 sucedem quando a partição é representada por grupos que compreendem um único objeto cada. Entretanto, este é um cenário bastante extremo e que possui pouca relevância prática. Na verdade, deseja-se que um índice de similaridade entre duas partições quaisquer resulte em valores próximos de zero, ou pelo menos em valores constantes, e, neste sentido, [27] propuseram uma versão modificada do índice Rand, originando o Ajusted Rand (R') expresso por $R' = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \left[1/\binom{n}{2}\right]\sum_{i=1}^{R}\binom{n_{i\cdot}}{2}\sum_{j=1}^{C}\binom{n_{\cdot j}}{2}}{(1/2)\left[\sum_{i=1}^{R}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{C}\binom{n_{\cdot j}}{2}\right] - \left[1/\binom{n}{2}\right]\sum_{i=1}^{R}\binom{n_{i\cdot}}{2}\sum_{j=1}^{C}\binom{n_{\cdot j}}{2}}$. O índice Ajusted Rand pode apresentar valores negativos, possui valor máximo igual a 1 e resulta em 0 quando o índice é igual ao seu valor esperado [25].

3 Agrupamento de objetos textuais

Um procedimento de agrupamento de objetos textuais tem como objetivo particionar automaticamente uma coleção de documentos em 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 $D = \{d_i\}, i = 1, \ldots, n$, (ii) um número desejado de grupos K e (iii) uma função objetivo que avalia a qualidade do agrupamento, deseja-se determinar uma associação $\gamma : D \rightarrow 1, \ldots, K$ 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 γ seja sobrejetiva, a fim de garantir que nenhum dos K grupos esteja vazio [24, 28].

3.1 Processo de agrupamento de objetos textuais

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 [29, 30]:

- 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 na representação dos 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 idéia 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. Por meio deste 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. A relação entre os documentos e os termos da coleção pode ainda ser retratada por meio de um vetor m-dimensional binário, $d_i^{'} = (w_{i1}^{'}, w_{i2}^{'}, \ldots, w_{im}^{'}), i = 1, \ldots, n$, no qual cada componente é determinada pela expressão $w_{ij}^{'} = 1$ se $t_j \in d_i^{'}$, ou pela expressão $w_{ij}^{'} = 0$ caso contrário. Observa-se que este modelo apresenta-se particularmente adequado quando do cálculo da dissimilaridade entre os documentos com o emprego da distância de Hamming, ao passo que o esquema *tf-idf* mostra-se mais apropriado quando da determinação da dessemelhança por intermédio da distâncias Euclidiana e do coseno, e dos coeficientes de Jaccard estendido e de correlação de Pearson;

- 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, da distância de Hamming, 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.

4 Algoritmo de agrupamento

Para execução dos ensaios que tiveram como propósito avaliar os índices de similaridade descritos na seção 2.1, o algoritmo *K-means* foi utilizado, haja vista que conforme [31], a sua simplicidade, a sua eficiência e os seus bons resultados experimentais contribuem para que este método seja um dos mais populares expedientes de particionamento, ainda que tenha sido inicialmente proposto há mais de 50 anos. Ademais, verifica-se que, a exemplo dos trabalhos de [32, 33, 34, 35, 36, 37], entre outros, esta estratégia têm sido assiduamente aplicada a problemas de classificação não supervisionada.

4.1 *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 $e_k^2 = \sum_{i=1}^{n_k} \parallel x_i^{(k)} - m^{(k)} \parallel^2$ [31, 18, 2, 1].

O objetivo do *K-means* consiste em minimizar a soma do erro quadrático sobre a partição constituída de K grupos, a qual será representada por $E_k^2 = \sum_{k=1}^{K} e_k^2$. 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 (com $E_k^2 = 0$ quando $K = n$), o seu valor é verdadeiramente minimizado apenas quando a quantidade de grupos permanece constante [31, 38, 18].

O método de agrupamento particional *K-means*, que foi proposto originalmente por [39], 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 [1, 31].

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 [31] e [40], 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 $w_{ij} = n_{ij} \log_2 \left(\frac{n}{n_j}\right)$ se a distância Euclidiana, a distância do coseno, o coeficiente de Jaccard estendido ou o coeficiente de correlação de Pearson estivesse sendo adotado, ou de acordo com a equação $w_{ij}^{'} = 1$ para $t_j \in d_i^{'}$ ou $w_{ij}^{'} = 0$ para $t_j \notin d_i^{'}$, se a distância de Hamming estivesse sendo utilizada;

- A média ou centróide de cada grupo foi determinada pela moda de cada componente dos vetores que pertenciam ao grupo, quando da adoção da distância de Hamming, ou conforme a expressão $m^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i^{(k)}$ se os demais índices de similaridade estivessem sendo empregados;

- A função objetivo a minimizar foi definida como o somatório da distância média dos documentos aos centróides dos grupos, e foi calculada de acordo com a equação $f = \left[\sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} d(m^{(k)}, d_i^{(k)})\right] \frac{1}{K}$ quando da utilização das distâncias Euclidiana ou do coseno, ou dos coeficientes de correlação de Pearson e de Jaccard estendido, ou ainda como o somatório do diâmetro dos grupos, representado pela expressão $f = \sum_{k=1}^{K} max\ d(d_i^{(k)}, d_j^{(k)})$, quando do emprego da distância de Hamming;

- 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.

5 Avaliação dos índices de similaridade

Com o objetivo de avaliar qual dentre os índices de proximidade representados pela distâncias Euclidiana, do coseno e de Hamming, e pelos coeficientes de Jaccard estendido e de correlação de Pearson, seria o mais adequado ao problema do agrupamento de documentos, um conjunto de experimentos foi realizado. Três coleções constituídas de objetos previamente classificados e com número de grupos conhecido foram selecionadas e os objetos constantes das mesmas foram submetidos ao método de agrupamento *K-means*, que foi codificado na linguagem de programação Microsoft Visual Basic .NET e executado em um micro-computador 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 site eletrônico 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 [41], 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 [42], 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 3 a seguir.

Tabela 3: Características das coleções de objetos textuais

Nome	Número de objetos	Número de dimensões	Número de grupos
Reuters 1	340	1.875	4
Reuters 2	403	1.895	8
Reuters 3	541	2.275	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

Tendo em conta que os objetos estavam antecipadamente categorizados e que o objetivo dos experimen-

tos era avaliar índices de proximidade, admitiu-se o emprego de critérios externos não dependentes da medida de similaridade utilizada pelo método de agrupamento, os quais foram representados pelos seguintes índices de validação de agrupamentos: Entropia, Pureza, Rand, Ajusted Rand, Fowlkes-Mallows e Γ Statistic. Para cada coleção de objetos e para cada índice de proximidade o algoritmo foi executado dez vezes, a fim de que as médias dos índices de validação, calculadas após o término do procedimento de agrupamento, pudessem ser comparadas. No intuito de auxiliar a aferição dos resultados, os índices de validação que admitiam valores 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 índice, 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 índice. Uma tabela de escores, que estabelecia valor 1 para o melhor resultado e também 1 para qualquer resultado distinto do melhor, desde que este estivesse até 5% aquém ou além do valor mais adequado, foi elaborada para cada conjunto de experimentos. Os escores alcançados pelos índices de proximidade ao se avaliar os índices de validação de agrupamentos foram somados, sendo o parecer mais favorável atribuído ao índice de proximidade que obtivesse a maior pontuação.

A tabela 4 demonstra os escores obtidos por cada um dos índices de similaridade confrontados, quando da aplicação do método *K-means* sobre as coleções de objetos determinadas. A análise dos valores expressos pela tabela de escores permite verificar que a distância do coseno obteve os resultados mais favoráveis. Em particular, observa-se que a distância do coseno computou 43 escores, enquanto que as distâncias Euclidiana e de Hamming, e os coeficientes de Jaccard estendido e de correlação de Pearson, apresentaram, nesta ordem, 3, 4, 31 e 6 escores. Estes achados sugerem portanto, que a distância do coseno seria, nesta avaliação, o índice de similaridade mais apropriado à classificação não supervisionada de objetos textuais. O número de escores auferido pelo coeficiente de Jaccard estendido, permite indicar que este índice de similaridade possui um comportamento semelhante ao apresentado pela distância do coseno, ainda que tenha obtido uma pontuação aproximadamente 20% inferior. Já os resultados verificados para as distâncias Euclidiana e de Hamming, e para o coeficiente de correlação de Pearson, denotam que estas medidas de similaridade são compatíveis entre si e que são menos convenientes ao particionamento de textos. A avaliação dos escores demonstrados na tabela 4 indica adicionalmente que a performance das medidas de similaridade utilizadas não foi influenciada pelo idioma, haja vista que para os textos em inglês o número de escores obtidos pelas distâncias Euclidiana, de Hamming e do coseno, e pelos coeficientes de Jaccard estendido e de correlação de Pearson, foram respectivamente 0, 1, 30, 17 e 3, ao passo que para os textos em português as pontuações dos mesmos índices foram, nesta ordem, 3, 3, 13, 14 e 3. Ou seja, a distância do coseno e o coeficiente de Jaccard estendido continuaram a denotar um comportamento mais conveniente, enquanto que as distância Euclidiana e de Hamming, e o coeficiente de correlação de Pearson mantiveram um desempenho congênere e menos adequado.

Tabela 4: Análise comparativa dos índices de similaridade quando da aplicação do método *K-means*. DE: Distância Euclidiana; DH: Distância de Hamming; DC: Distância do coseno; CJ: Coeficiente de Jaccard estendido; CP: Coeficiente de correlação de Pearson; EE: Escores da distância Euclidiana; EH: Escores da distância de Hamming; EC: Escores da distância do coseno; EJ: Escores do coeficiente de Jaccard estendido; EP: Escores do coeficiente de correlação de Pearson

Índice de validação	Conjunto de dados	DE	DH	DC	CJ	CP	EE	EH	EC	EJ	EP
Entropia	Reuters 1	0,9310	0,9679	**0,3013**	0,3691	0,5584	0	0	1	0	0
	Reuters 2	0,8775	0,8892	**0,4729**	**0,4902**	0,6748	0	0	1	1	0
	Reuters 3	0,8342	0,8624	**0,4792**	**0,4741**	0,6348	0	0	1	1	0
	Newsgroups 1	0,9950	0,8847	0,7350	**0,5590**	0,9292	0	0	0	1	0
	Newsgroups 2	0,9710	0,9847	**0,6950**	0,7876	0,8838	0	0	1	0	0
	Newsgroups 3	0,9795	0,9837	**0,6407**	0,7663	0,8137	0	0	1	0	0
	Jornal 1	0,9523	0,9526	**0,7610**	**0,7560**	0,8922	0	0	1	1	0
	Jornal 2	0,9230	0,9217	**0,6754**	**0,6574**	0,7708	0	0	1	1	0
	Jornal 3	0,9335	0,9393	**0,6288**	**0,6036**	0,6982	0	0	1	1	0
Pureza	Reuters 1	0,3677	0,3388	**0,8594**	**0,8294**	0,6862	0	0	1	1	0
	Reuters 2	0,2697	0,2489	**0,6538**	**0,6347**	0,4918	0	0	1	1	0
	Reuters 3	0,2647	0,2357	**0,5691**	**0,5869**	0,4496	0	0	1	1	0
	Newsgroups 1	0,5050	0,6060	0,7220	**0,8275**	0,6495	0	0	0	1	0
	Newsgroups 2	0,2980	0,2655	**0,5697**	0,5085	0,4495	0	0	1	0	0
	Newsgroups 3	0,1905	0,1855	**0,5537**	0,4458	0,4240	0	0	1	0	0
	Jornal 1	0,3399	0,3404	**0,5488**	**0,5390**	0,4540	0	0	1	1	0
	Jornal 2	0,2013	0,2013	0,4724	**0,5048**	0,4099	0	0	0	1	0
	Jornal 3	0,1405	0,1408	**0,4693**	**0,4860**	0,4168	0	0	1	1	0
Rand	Reuters 1	0,3322	0,2827	**0,8991**	**0,8588**	0,7764	0	0	1	1	0
	Reuters 2	0,3513	0,3569	**0,8543**	**0,8446**	0,8103	0	0	1	1	0
	Reuters 3	0,4370	0,3943	**0,8673**	**0,8714**	**0,8456**	0	0	1	1	1
	Newsgroups 1	0,4975	0,5654	0,6227	**0,7519**	0,5438	0	0	0	1	0
	Newsgroups 2	0,3509	0,2645	**0,6713**	0,6304	**0,6498**	0	0	1	0	1
	Newsgroups 3	0,2282	0,2006	**0,7648**	0,6821	**0,7398**	0	0	1	0	1
	Jornal 1	0,2819	0,2822	**0,6489**	**0,6459**	**0,6358**	0	0	1	1	1
	Jornal 2	0,1789	0,1929	**0,7758**	**0,7955**	**0,7875**	0	0	1	1	1
	Jornal 3	0,1290	0,1351	**0,8523**	**0,8566**	**0,8557**	0	0	1	1	1
Ajusted Rand	Reuters 1	0,0364	0,0048	**0,7436**	0,6485	0,4264	0	0	1	0	0
	Reuters 2	0,0279	0,0177	**0,4214**	0,3847	0,2428	0	0	1	0	0
	Reuters 3	0,0280	0,0242	**0,3314**	**0,3302**	0,1984	0	0	1	1	0
	Newsgroups 1	0,0000	0,1342	0,2468	**0,5045**	0,0878	0	0	0	1	0
	Newsgroups 2	0,0105	0,0003	**0,2569**	0,1877	0,1021	0	0	1	0	0
	Newsgroups 3	0,0025	0,0012	**0,2846**	0,1616	0,1471	0	0	1	0	0
	Jornal 1	0,0019	0,0023	**0,1817**	0,1635	0,0682	0	0	1	0	0
	Jornal 2	0,0008	0,0022	0,1690	**0,2224**	0,1297	0	0	0	1	0
	Jornal 3	0,0005	0,0003	**0,2296**	**0,2344**	0,1666	0	0	1	1	0
Fowlkes-Mallows	Reuters 1	0,5017	0,4976	**0,8137**	0,7471	0,5787	0	0	1	0	0
	Reuters 2	0,3384	0,3285	**0,5078**	0,4767	0,3542	0	0	1	0	0
	Reuters 3	0,2799	0,2840	**0,4068**	**0,4025**	0,2851	0	0	1	1	0
	Newsgroups 1	0,7018	0,7072	0,6889	**0,7793**	0,5503	0	0	0	1	0
	Newsgroups 2	0,4524	**0,4903**	**0,4922**	0,4480	0,3396	0	1	1	0	0
	Newsgroups 3	0,3883	0,3970	**0,4334**	0,3597	0,3065	0	0	1	0	0
	Jornal 1	**0,5105**	**0,5108**	0,4316	0,4146	0,3163	1	1	0	0	0
	Jornal 2	**0,3766**	**0,3744**	0,3024	0,3435	0,2535	1	1	0	0	0
	Jornal 3	**0,3042**	**0,3026**	**0,3129**	**0,3149**	0,2464	1	1	1	1	0
Γ Statistic	Reuters 1	0,0919	0,0432	**0,9182**	0,8056	0,5369	0	0	1	0	0
	Reuters 2	0,1033	0,0764	**0,8379**	0,7660	0,4876	0	0	1	0	0
	Reuters 3	0,1494	0,1232	**0,8540**	**0,8487**	0,5203	0	0	1	1	0
	Newsgroups 1	0,0014	0,1673	0,3138	**0,6260**	0,1098	0	0	0	1	0
	Newsgroups 2	0,0337	0,0020	**0,5584**	0,4002	0,2121	0	0	1	0	0
	Newsgroups 3	0,0127	0,0074	**0,6599**	0,3888	0,3352	0	0	1	0	0
	Jornal 1	0,0784	0,0840	**0,5375**	0,4866	0,2331	0	0	1	0	0
	Jornal 2	0,0491	0,0649	0,5522	**0,7143**	0,4295	0	0	0	1	0
	Jornal 3	0,0335	0,0280	**0,7729**	**0,7866**	0,5631	0	0	1	1	0
Total							**3**	**4**	**43**	**31**	**6**

No intuito de avaliar a variabilidade dos resultados obtidos pelos índices de similaridades mais apropriados, neste estudo representados pela distância do coseno e pelo coeficiente de Jaccard estendido, o desvio padrão e o desvio absoluto médio foram retratados por intermédio de duas novas tabelas de escores, que foram elaboradas de modo analógo ao adotado na comparação direta entre os índices de similaridade. O primeiro conjunto de resultados, denotado por intermédio da tabela 5, apresenta o desvio padrão dos índices de validação de agrupamentos, enquanto que a tabela 6 expressa para os mesmos índices de proximidade e para as mesmas coleções de textos, os desvios absolutos médios dos índices de validação de particionamento. Os escores auferidos pela distância do coseno e pelo coeficiente de Jaccard estendido permitem indicar que a distância do coseno possui, quanto à dispersão dos resultados em relação à média, um comportamento mais favorável, haja vista que para o desvio padrão a quantidade de escores computada pela distância do coseno e pelo coeficiente de Jaccard estendido, foram respectivamente, 6 e 1, ao passo que para o desvio absoluto médio os escores foram nesta ordem 6 e 0.

Tabela 5: Desvio padrão dos índices de validação de agrupamentos Entropia, Pureza, Rand, Ajusted Rand, Fowlkes-Mallows e Γ Statistic, quando da aplicação do método *K-means* sobre as coleções de objetos textuais, com o emprego dos índices de similaridade distância do coseno e coeficiente de Jaccard estendido. DC: Distância do coseno; CJ: Coeficiente de Jaccard estendido; EC: Escores da distância do coseno; EJ: Escores do coeficiente de Jaccard estendido

Método	Índice de validação	DC	CJ	EC	EJ
K-means	Entropia	**0,1617**	0,1739	1	0
	Pureza	**0,1309**	0,1566	1	0
	Rand	**0,1100**	**0,1141**	1	1
	Ajusted Rand	**0,1911**	0,2032	1	0
	Fowlkes-Mallows	**0,1629**	0,1716	1	0
	Γ Statistic	**0,1901**	0,2017	1	0
Total				**6**	**1**

Tabela 6: Desvio absoluto médio dos índices de validação de agrupamentos Entropia, Pureza, Rand, Ajusted Rand, Fowlkes-Mallows e Γ Statistic, quando da aplicação do métodos *K-means* sobre as coleções de objetos textuais, com o emprego dos índices de similaridade distância do coseno e coeficiente de Jaccard estendido. DC: Distância do coseno; CJ: Coeficiente de Jaccard estendido; EC: Escores da distância do coseno; EJ: Escores do coeficiente de Jaccard estendido

Método	Índice de validação	DC	CJ	EC	EJ
K-means	Entropia	**0,1297**	0,1440	1	0
	Pureza	**0,0969**	0,1211	1	0
	Rand	**0,0923**	0,0971	1	0
	Ajusted Rand	**0,1389**	0,1586	1	0
	Fowlkes-Mallows	**0,1277**	0,1353	1	0
	Γ Statistic	**0,1381**	0,1575	1	0
Total				**6**	**0**

Uma análise complementar que teve como propósito avaliar o tempo de execução do método de particionamento, quando da utilização dos índices de similaridade distância do coseno e coeficiente de Jaccard estendido, conduziu à elaboração da tabela 7, confeccionada de modo semelhante às tabelas de pontuação anteriormente mencionadas. O número de escores obtidos pelos índices de similaridade denota que a distância do coseno possui uma performance invariavelmente superior à do coeficiente de Jaccard estendido, por ocasião do emprego do método *K-means*. Em particular, verifica-se que o número de escores auferido pela a distância do coseno foi nove, enquanto que a quantidade de escores computada pelo coeficiente de Jaccard estendido foi zero. Considerando-se que o algoritmo *K-means* emprega a média dos objetos como centro dos grupos, e que demanda que as similaridades entre os objetos e os centros dos grupos até então determinados tenham que ser continuamente calculadas, sugere-se que o menor tempo de execução observado quando do emprego da distância do coseno é resultante da menor esforço de processamento necessário à determinação da distância entre os objetos, haja vista que a equação que expressa a

distância do coseno é menos extensa do que a igualdade que representa o coeficiente de Jaccard estendido.

Tabela 7: Tempo médio de execução, em segundos, do método *K-means* sobre as coleções de objetos textuais, com o emprego dos índices de similaridade distância do coseno e coeficiente de Jaccard estendido. DC: Tempo de execução com o uso da distância do coseno; CJ: Tempo de execução com o uso do coeficiente de Jaccard estendido; EC: Escores da distância do coseno; EJ: Escores do coeficiente de Jaccard estendido

Método	Conjunto de dados	DC	CJ	EC	EJ
K-means	Reuters 1	**257**	378	1	0
	Reuters 2	**592**	783	1	0
	Reuters 3	**1.483**	1.871	1	0
	Newsgroups 1	**91**	121	1	0
	Newsgroups 2	**868**	913	1	0
	Newsgroups 3	**2.630**	2.796	1	0
	Jornal 1	**169**	222	1	0
	Jornal 2	**1.148**	1.473	1	0
	Jornal 3	**5.216**	5.982	1	0
Total				**9**	**0**

6 Conclusão

Diante do elevado e crescente volume 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.

Ante o progressivo acúmulo de dados textuais em formato eletrônico, diversas técnicas de mineração têm sido propostas no intuito de endereçar as dificuldades decorrentes da disponibilidade deste elevado volume de informações, e, dentre elas, destacam-se àquelas relacionadas às atividades de análise de agrupamentos, que têm como propósito fundamental particionar os objetos em grupos distintos de forma que a distância entre os elementos que fazem de um mesmo grupo seja minimizada, ao mesmo tempo que a distância entre os objetos que pertencem a grupos distintos seja maximizada. Tendo em conta que o índice de similaridade empregado na determinação da semelhança ou distância entre os objetos exerce significante influência sobre os resultados obtidos pelos algoritmos de agrupamento, este estudo teve como propósito avaliar cinco índices de similaridade distintos. Por intermédio de experimentos realizados sobre coleções de objetos heterogêneas, utilizando o algoritmo *K-means*, e sob a perspectiva de índices de validação de partições, identificou-se a distância do coseno e o coeficiente de Jaccard estendido como os índices de similaridade mais adequados à determinação da dessemelhança entre os vetores utilizados para retratar os documentos. Verificou-se adicionalmente que não houve dependência entre o idioma dos textos utilizados nos ensaios e os resultados auferidos pelos índices de similaridades mais apropriados, haja vista que tanto para os documentos em português quanto para os documentos em inglês a distância do coseno e o coeficiente de Jaccard estendido obtiveram os maiores escores. Ademais, observou-se que sob a perspectiva da variabilidade dos resultados e dos tempos de execução do algoritmo de particionamento, a distância do coseno obteve um melhor comportamento quando diretamente confrontada com o coeficiente de Jaccard estendido, sugerindo por consequência que o primeiro seria, dentre os avaliados, o índice de similaridade mais apropriado à classificação não supervisionada de objetos textuais.

Referências

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

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

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

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

[5] ALIGULIYEV, R. M. Clustering of document collection a weighting approach. *Expert Systems with Applications*, v. 36, n. 4, p. 7904–7916, 2009.

[6] A, K. K.; PHILIP, M.; LUBNA, K. Comparative analysis of similarity measures in document clustering. In: *Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on*. Chennai, India: IEEE, 2013. p. 857–860.

[7] SRUTHI, K.; VENKATESHWAR, B. Document clustering on various similarity measures. *International Journal of Advanced Research in Computer Science and Software Engineering*, v. 3, n. 8, p. 1269–1273, 2013.

[8] NAGWANI, N. K. A comment on "a similarity measure for text classification and clustering". *IEEE Transactions on Knowledge and Data Engineering*, v. 27, n. 9, p. 2589–2590, 2015.

[9] WANG, C. et al. Knowsim: A document similarity measure on structured heterogeneous information networks. In: *2015 IEEE International Conference on Data Mining*. Atlantic City, New Jersey: IEEE, 2015. p. 1015–1020.

[10] NALAWADE, R.; SAMAL, A.; AVHAD, K. Improved similarity measure for text classification and clustering. *International Research Journal of Engineering and Technology*, v. 3, n. 5, p. 214–219, 2016.

[11] POTDAR, D. S.; PATTEWAR, T. M. A novel similarity measure technique for clustering using multiple viewpoint based method. In: *2016 10th International Conference on Intelligent Systems and Control (ISCO)*. Coimbatore, India: IEEE, 2016. p. 1–4.

[12] THOMAS, A. M.; RESMIPRIYA, M. An efficient text classification scheme using clustering. *Procedia Technology*, v. 24, n. Supplement C, p. 1220–1225, 2016.

[13] HEIDARIAN, A.; DINNEEN, M. J. A hybrid geometric approach for measuring similarity level among documents and document clustering. In: *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*. Oxford, UK: IEEE, 2016. p. 142–151.

[14] AISHWARYA, M. L.; SELVI, K. An intelligent similarity measure for effective text document clustering. In: *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*. Kovilpatti, India: IEEE, 2016. p. 1–5.

[15] SOHANGIR, S.; WANG, D. Document understanding using improved sqrt-cosine similarity. In: *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. San Diego, California: IEEE, 2017. p. 278–279.

[16] JAGATHEESHKUMAR, G.; BRUNDA, S. S. An analysis of efficient clustering methods for estimates similarity measures. In: *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*. Coimbatore, India: IEEE, 2017. p. 1–3.

[17] XU, R.; WUNSCH, D. C. *Clustering*. Piscataway, New Jersey: IEEE Press, 2009.

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

[19] EVERITT, B. S. et al. *Cluster Analysis*. London: John Wiley & Sons, Ltd, 2011.

[20] GASIENIEC, L.; JANSSON, J.; LINGAS, A. Approximation algorithms for hamming clustering problems. *Journal of Discrete Algorithms*, n. 2, p. 289–301, 2004.

[21] SMET, W. D.; MOENS, M.-F. Representations for multi–document event clustering. *Data Mining and Knowledge Discovery*, v. 26, n. 3, p. 333–558, 2013.

[22] SANDHYA, N.; GOVARDHAN, A. Analysis of similarity measures with wordnet based text document clustering. In: ____. *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 703–714.

[23] CONRAD, J. G. et al. Effective document clustering for large heterogeneous law firm collections. In: *Proceedings of the 10th International Conference on Artificial Intelligence and Law*. Bologna, Italy: ACM, 2005. p. 177–187.

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

[25] VINH, N. X.; EPPS, J.; BAILEY, J. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In: *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal, Canada: ACM, 2009. p. 1073–1080.

[26] BOCK, H.-H.; POLASEK, W. *Data Analysis and Information Systems: Statistical and Conceptual Approaches*. Berlin: Springer, 1996.

[27] HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, v. 2, n. 1, p. 193–218, 1985.

[28] KALOGERATOS, A.; LIKAS, A. Text document clustering using global term context vectors. *Knowledge and Information Systems*, v. 31, n. 3, p. 455–474, 2012.

[29] KAROL, S.; MAGNAT, V. Evaluation of text document clustering approach based on particle swarm optimization. *Central European Journal of Computer Science*, v. 2, n. 3, p. 69–90, 2013.

[30] TSENG, Y.-H. Generic title labeling for clustered documents. *Expert Systems with Applications*, v. 37, n. 3, p. 2247–2254, 2010.

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

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

[33] 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. p. 265–268.

[34] KANT, S.; ANSARI, I. A. An improved k–means clustering with atkinson index to classify liver patient dataset. *International Journal of System Assurance Engineering and Management*, v. 7, n. 1, p. 222–228, 2016.

[35] VINUÉ, G.; SIMÓ, A.; ALEMANY, S. The k–means algorithm for 3d shapes with an application to apparel design. *Advances in Data Analysis and Classification*, v. 10, n. 1, p. 103–132, 2016.

[36] GANESH, M.; NARESH, M.; ARVIND, C. Mri brain image segmentation using enhanced adaptive fuzzy k-means algorithm. *Intelligent Automation & Soft Computing*, v. 23, n. 2, p. 325–330, 2017.

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

[38] DRINEAS, P. et al. Clustering large graphs via the singular value decomposition. *Machine Learning*, v. 56, n. 1–3, p. 9–33, 2004.

[39] 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. p. 281–297.

[40] FORSATI, R. et al. Efficient stochastic algorithms for document clustering. *Information Sciences*, v. 220, p. 269–291, 2013.

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

[42] 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. p. 186–193.