ALGORITMOS *FUZZY* APLICADOS EM TEXTOS DE NUTRIGENÔMICA

Carla Cristina Passos Cruz
Universidade do Estado do Rio de Janeiro – UERJ
Instituto de Matemática e Estatística – IME/UERJ
Rua São Francisco Xavier, 524, Pavilhão João Lyra Filho, 6º andar, Maracanã – RJ
e-mail: carlapassos2889@gmail.com

Regina Serrão Lanzillotti
Universidade do Estado do Rio de Janeiro – UERJ
Instituto de Matemática e Estatística – IME/UERJ
Rua São Francisco Xavier, 524, Pavilhão João Lyra Filho, 6º andar, Maracanã – RJ
e-mail: reginalanzillotti@gmail.com

Haydée Serrão Lanzillotti
Universidade do Estado do Rio de Janeiro – UERJ
Instituto de Nutrição – NUT/UERJ
Rua São Francisco Xavier, 524, Pavilhão João Lyra Filho, 12º andar, Maracanã – RJ
e-mail: haydeelan@gmail.com

RESUMO

Este artigo apresenta a mineração de três textos sobre o tema Nutrigenômica, cujo objetivo foi confrontar os resultados referentes aos agrupamentos criados pelos métodos não-hierárquicos *fuzzy* em Mineração de Textos. Foram aplicados dois algoritmos de agrupamento *fuzzy C-Means e fuzzy C-Medoids*, sendo que o primeiro se mostrou mais eficiente para discriminação de termos genômicos. A visualização gráfica permitiu a interpretação mais adequada para sumarizar os resultados, inclusive destacando o agrupamento que configura um cenário próximo da nitidez sob a ótica semântica. A mineração de texto pode auxiliar no desempenho de aceleração da tarefa da busca de informações em acervos vastos e dispersos. O uso de algoritmos na Mineração de Texto otimiza a busca em função das palavras-chave geradas por eles. No presente estudo o confronto dos algoritmos *fuzzy C-Means* e *fuzzy C-Medoids*, mostrou que o primeiro agregou os termos sob a ótica semântica com efetividade.

Palavra-chave: algoritmos *fuzzy*; mineração de texto; Nutrigenômica.

ABSTRACT

This article presents a search tool for three texts on the Nutrigenomic theme, which objective was to compare the results by grouping through fuzzy non-hierarchical methods in Text Mining. Fuzzy C-Means and fuzzy C-Medoids presents to be the most efficient for discriminating genomic terms. The graphic preview allowed a more complete interpretation for the summary. of the results, including the grouping that configures a next principle of sharpness from the semantic point of view. Text mining is a powerful source to assist in

accelerating the performance of the vast and dispersed information seeking task. The use of text mining algorithms optimizes the search for the key functions generated by them. This article presents the parallel of fuzzy C-Means and fuzzy C-Medoids algorithms, results on the first aggregating terms under the semantic optics with more efficiency.

Keywords: fuzzy algorithms; text mining; Nutrigenomics.

Como Citar:
CRUZ, Carla Cristina Passos; LANZILLOTTI, Regina Serrão; LANZILLOTTI, Haydée Serrão. Algoritmos *Fuzzy* Aplicados em Textos de Nutrigenômica. *In*: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA, 19, 2019, Rio de Janeiro, RJ. **Anais** […]. Rio de Janeiro: Centro de Análises de Sistemas Navais, 2019.

1. INTRODUÇÃO

No mundo atual, é inegável que a tecnologia esteja presente em todas as ações do cotidiano, sendo que a Inteligência Artificial propicia a interação entre o homem e a máquina, o que torna viável o tratamento na Ciência de Dados. A associação homem/máquina envolve interações tais como postagens nas redes sociais, pagamento de contas via celular, compras em *sites*, cadastramento de clientes, dentre outros. De acordo com Hilbert e Lopez [1], 90% dos dados do mundo criados nos últimos anos foram decorrentes a adesão das grandes empresas à *internet*. Em muitas situações, a mineração de dados se faz necessária [2], devido à necessidade de armazenamento de forma adequada para o uso em algoritmos. Deste modo, torna-se um desafio minerar, limpar, organizar, correlacionar, vincular e transformar esses dados em informações relevantes, uma vez que Tessarolo e Magalhães [3] apontaram que o índice correspondente à mineração é inferior a 10%.

No entanto, analisar grandes volumes de dados torna-se inviável se não houver um suporte computacional, pois há necessidade de se aplicar técnicas de forma a extrair informações. De acordo com Maimon e Rokach [4], o processo de *KDD - Knowledge Discovery in Databases* pode ser entendido como um conjunto de etapas que se iniciam em função da definição de uma meta. O *KDT – Knowledge Discovery in Text* se realiza a mineração de textos, processo em que Tan [5] é citado em diversas publicações e indica que 80% das informações vigentes são não-estruturadas. Desta forma, surge à necessidade da aplicação de técnicas que possibilitem a extração, análise e resultados, pois as análises textuais podem ser úteis em diversas situações, uma vez que propicia captar a opinião dos usuários sobre uma determinada temática ou ainda realizar uma análise da reação das pessoas. Krassmann *et al.* [6] usaram a mineração de texto para analisar as tendências em trabalhos acadêmicos sobre mundos virtuais na educação com o método de pesquisa em que selecionou cinco termos para compor um *string* e mostraram que a técnica de mineração de texto é realizada de forma automática, rápida e eficiente de analisar um volume expressivo de dados sem que haja a necessidade de leitura analítica, o que exigiria tempo.

Este trabalho tem como objetivo confrontar os resultados referentes aos agrupamentos criados pelos métodos não-hierárquicos *fuzzy* em Mineração de Textos, *C-means* e *C-medoids*. O artigo está dividido em cinco seções, além desta Introdução. A seção dois descreve os conceitos referentes a Mineração de Texto. A seção três explicita os métodos *fuzzy C-Means* e *fuzzy C-Medoids*. A seção quatro traz o conjunto de resultados a partir da aplicação dos algoritmos e o confronto entre os dois métodos. Finalmente, a seção cinco traz as conclusões e as considerações finais.

2. MINERAÇÃO DE TEXTO

A Mineração de Texto (MT) é um processo de Descoberta de Conhecimento que utiliza técnicas de análise e extração de dados a partir de diferentes tipos de textos, além de envolver a aplicação de algoritmos computacionais que os processam, identificando informações úteis e implícitas que não poderiam ser recuperadas por métodos tradicionais de consulta, uma vez que a informação obtida encontra-se em formato não-estruturado [7].

A organização de um conjunto de documentos ou termos em grupos tem sido adaptado as principais áreas de conhecimento que compõem e contribuem com a Mineração de Texto. Dentre as quais, tem se utilizado a Recuperação da Informação (RI), o Aprendizado de Máquina (AM) [8], Processamento de Linguagem Natural (PLN), Estatística Inferencial [9], a Inteligência Computacional (IC), a Ciência Cognitiva, a Mineração de Dados e *Web Mining* [10]. Objetiva-se organizar um conjunto de documentos ou termos em grupos, de forma que os mesmos tenham conteúdo similar, isto é, tratem de assuntos parecidos.

O agrupamento é um método utilizado para identificar relacionamentos entre objetos, facilitando a identificação de classes. No caso de textos, o agrupamento identifica conteúdos similares, caso não se tenha definição dos assuntos tratados em cada texto e se deseja separá-los por assunto [11]. Patil e Dongre [12] tratam como classificação *hard*, quando os dados pertencem a apenas um agrupamento, e em *soft*, caso contrário. Assim, a especificação *soft* leva ao agrupamento *fuzzy*, pois cada ponto de dados é associado a uma função que expressa o grau de pertinência dos dados de forma individualizada. A precisão é obtida na etapa de Recuperação da Informação, utilizando-se como medida o *fuzzy clustering*, processo de particionamento não-hierárquico no qual se busca a divisão do conjunto de entidades em um número de grupos homogêneos.

De acordo com Goularte [13], a informação na forma textual é representada por termos linguísticos que trazem consigo a incerteza, que está envolvida na resolução de um problema que pode ser decorrente de alguma informação deficiente ou porque existe mais de uma solução. Dessa forma, modelagem *fuzzy* é uma importante abordagem amplamente recomendada para aplicações cujo domínio esteja caracterizado por incerteza ou imprecisão da informação [14].

Há diferença entre o agrupamento não-hierárquico tradicional e o *fuzzy:* enquanto no primeiro, o objeto pertence a um único conglomerado, no segundo um objeto pode pertencer simultaneamente a mais de um grupo, com diferentes graus de pertinência [15]. Nesta proposta, se utilizam os algoritmos *fuzzy C-means* e *fuzzy C-medoids* para agrupar os termos de três documentos inerentes ao contexto da genômica nutricional**,** estudo do impacto de nutrientes na expressão gênica, o que permite conhecer o mecanismo de ação das substâncias biologicamente ativas, contidas nos alimentos, e seus efeitos benéficos no organismo.

3. MÉTODOS DE AGRUPAMENTO *FUZZY*

O agrupamento consiste em dividir uma amostra ou população em grupos, de acordo com as características medidas, de forma que elementos pertencentes a um mesmo grupo sejam homogêneos (semelhantes) entre si, e elementos em grupos diferentes sejam heterogêneos com relação às características [9]. No entanto, um dos maiores problemas é a identificação de termos quando possa pertencer a mais de um grupo simultaneamente, o que conduz aos métodos *fuzzy* que estabelecem o grau de pertinência deste termo em cada um desses conjuntos. Para aplicação da técnica de agrupamento, Rodrigues [16] sugere as etapas:

1) Análise dos termos (variáveis) a serem agrupadas, identificação de *outliers* e padronização;
2) Seleção da medida de similaridade entre cada par de observações;

3) Seleção do algoritmo de agrupamento: método hierárquico ou não-hierárquico;
4) Definição da quantidade de conglomerados formados;
5) Interpretação e validação dos conglomerados.

Neste trabalho cada termo foi identificado segundo a frequência relativa e adotado um corte de um limiar percentual para estabelecer os termos que deveriam constar no desenvolvimento dos métodos *fuzzy*. As frequências relativas foram redimensionadas segundo escores ranqueados, sendo que a frequência máxima correspondeu ao maior escore, enquanto a mínima o menor. Os escores ranqueados permitiram obter uma matriz de correlação segundo *Spearman* [17,18] para estabelecer o número de agrupamentos a serem inseridos nos métodos *fuzzy C-Means* e *fuzzy C-Medoids*, e o confronto dos resultados indicou o melhor deles.

3.1 *FUZZY C-MEANS*

O primeiro algoritmo a ser utilizado será o *fuzzy C-Means* (FCM) desenvolvido por Dunn [19] e melhorado por Bezdek [20]. Cada conglomerado tem um agrupamento central que representa um objeto típico e o valor de associação é expresso segundo pertinência de um objeto que está próximo ao centro do conglomerado. Desta forma, se o valor da associação for alto, o elemento será semelhante ao elemento ao centro do conglomerado e, caso seja baixo, o elemento tem pouca semelhança com os elementos do conglomerado [21]. Mingoti [9] aponta para uma pré-especificação do número de grupos. O algoritmo FCM agrupa termos inerentes a genômica nutricional em grupos de cenários que permitem identificar padrões textuais da ação das substâncias biologicamente ativas, contidas nos alimentos e seus efeitos benéficos no organismo.

No processo de partição do FCM, dado um conjunto de dados $X = \{x_i \mid i = 1, \cdots, k\}$, onde x_k é um vetor de características $x_k = \{x_{k_1}, x_{k_2}, ..., x_{k_p}\} \in \Re^p$ para todo $k \in \{1,2,..n\}$ e sendo $\Re^p$ o espaço p-dimensional. O problema da aglomeração *fuzzy* é encontrar uma pseudopartição que represente a estrutura dos dados da melhor forma possível, em outras palavras, se busca minimizar a função objetivo dada por:

$$J(V;X) = \sum_{i=1}^{n}\sum_{j=1}^{g} u_{ij}^{m} d\left(x_i, v_j\right), \tag{1}$$

onde:

- n é a quantidade de elementos;
- g é o número de grupos pré-determinado;
- m é o parâmetro de *fuzzificação*, cujo valor sugerido por Pal e Bezdek [22], sendo o ponto médio $m = 2$, a escolha mais usual na literatura;
- v_j é o centro do agrupamento *fuzzy*;
- $d\left(x_j, v_i\right)$ é a distância euclidiana entre o elemento da amostra e o centro do agrupamento.

Uma pseudopartição *fuzzy* de X é uma família de k subconjuntos *fuzzy* de X, denotada por:

$$U = \left(u_1, u_2, ..., u_k\right) \tag{2}$$

chamada de pertinências *fuzzy*, obtidas pelo critério de normalização que satisfaz a condição:

$$\sum_{k=1}^{n_k}\sum_{j=1}^{p} u_{kj} = 1. \tag{3}$$

Dessa forma, o fluxograma apresenta as etapas do algoritmo *fuzzy C-Means* no Quadro 1, segundo os passos:

Quadro 1 – Etapas do Método *Fuzzy C-Means*

Etapas do Método Fuzzy C-Means	
Entrada: *Z = matriz com os escores ranqueados padronizados em unidades do desvio-padrão correspondentes a* n_{kj} *termos; m= 2, parâmetro de fuzzificação; k, quantidade de agrupamentos obtidos pelo gráfico das correlações;* $n = \sum_{k=1}^{n_k}\sum_{j=1}^{p} n_{kj}$ *quantidade total de termos da matriz Z.*	
Passo1	Obtenha as pertinências iniciais $u_{(0)_{kj}}$ geradas de forma aleatória para cada agrupamento, adotando-se o critério de normalização $\sum_{k=1}^{n_k}\sum_{j=1}^{p} u_{kj} = 1$;
Passo2	Obtenha o centro inicial $v_{(0)_k}$ para cada agrupamento gerado segundo uma média ponderada em função dos valores padronizados correspondentes aos termos, cujas pertinências correspondem aos ponderadores: $v_{0_k} = \frac{\sum_{j=1}^{p} u_{kj}^m Z_{kj}}{\sum_{j=1}^{p} u_{kj}^m}$;
Passo3	Calcule a distância euclidiana entre os elementos e o centro do agrupamento: $d_{0_k}(z_{kj}, v_{0k}) = \|z_{kj} - v_{0k}\| = \sqrt{\sum_{j=1}^{p}(z_{kj} - v_{0k})^2}$;
Passo4	Calcule a função objetivo inicial $J_0(Z;V) = \sum_{j=1}^{p}\sum_{k=1}^{n_k} u_{ij}^m d(z_{kj}, v_{0k})$;
Passo5	Atualize as pertinências u_{kj} de forma que, para cada $z_{kj} \in Z_k$, calcule o grau das novas pertinências: $u_{kj} = \frac{\sum_{k=1}^{n}\left(\frac{1}{d(z_{kj}, v_{0k})}\right)^{\frac{2}{m-1}}}{\sum_{k=1}^{n} u_{kj}^m}$;
Passo6	Volte para o **Passo2** e obtenha os centroides atualizados, $v_j = (v_1, v_2, ..., v_p) \in \Re^{cp}, v_j \in \Re^p \forall j$;
Passo7	Volte para o **Passo3**;
Passo8	Volte para o **Passo4. SE** a função objetivo estiver minimizada; **FIM;** **SENÃO** Volte para o **Passo5**.
Saída: *Agrupamentos segundo temas com respectivos termos inerentes a genômica nutricional.*	

Fonte: A autora, 2019.

3.2 *FUZZY C-MEDOIDS*

O segundo algoritmo a ser abordado neste artigo é o *C-Medoids*, uma versão *Fuzzy* criada por Krishnapuram *et al.* [24] com base nos algoritmos de particionamento rígido *K-*

Medoids [25] e o método CLARA [26]. O *k-medoids* não utiliza a média como centro do grupo, mas seleciona um objeto como se fosse o centro do próprio grupo. Esta opção se baseia em encontrar *k* objetos representativos, os chamados *medoids*, que em função do conjunto de termos vem a minimizar a soma das dissimilaridades dentro do agrupamento [15].

O desenvolvimento do método calcula os *medoids* (termos) iniciais para propiciar a formação dos grupos que indicarão os cenários dos genomas nutricionais. Cabe identificar se os termos não considerados *medoids* pertencem a um agrupamento liderado por um termo (*medoid*). Caso não pertença a um *medoid* característico, o algoritmo indicará a sua alocação com o respectivo grau de pertinência, mas o termo também poderá pertencer a um dos *medoids* mais próximos.

O *Fuzzy C-Medoids* (FCMdd), tem como base uma função objetivo para agrupamento relacional *fuzzy* a partir da identificação de *k* objetos representativos de cada agrupamento (*medoids*) que minimizem a dissimilaridade dentro de cada grupo. Seja $X = \{X_i \mid i = 1, \cdots, n\}$ conjunto com *n* observações, d_{ij} corresponde à distância entre os objetos de X [15]. A função de pertinência sugerida Krishnapuram, Joshi e Yi [27] apresentaram a seguinte expressão:

$$u_{ij} = \exp\left(-\frac{d(x_j - v_i)}{\eta_i}\right), \tag{4}$$

onde o parâmetro η_i necessita ser pré-fixado e pode ser estimado da distância euclidiana do conjunto X (KRISHNAPURAM; KELLER, 1996). O algoritmo *Fuzzy C-Medoids* minimiza:

$$J(V;X) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^{m} d\left(x_j, v_i\right) \tag{5}$$

onde a minimização é realizada para todo V em X_k. A função u_{ij}^{m} representa os elementos da matriz de pertinência *fuzzy* de X_j no agrupamento *i*, tendo *m* como parâmetro de *desfuzzificação* com valor no intervalo [1,∞) sendo adotado *m = 2* [28,15] assim como no FCM. Quando a função objetivo é minimizada, os *medoids* correspondentes a solução otimizada geram a partição de possibilidade *fuzzy* via pertinências.

Quadro 2 – Etapas do Método *Fuzzy C-Medoids* (continua)

Etapas do Método Fuzzy C-Medoids	
Entrada: *Z = matriz com os escores ranqueados padronizados em unidades do desvio-padrão correspondentes a* n_{kj} *termos; m= 2, parâmetro de fuzzificação; n =* $\sum_{k=1}^{n_k}\sum_{j=1}^{p} n_{kj}$ *quantidade total de termos da matriz Z.*	
Passo1	Fixe a quantidade de agrupamentos obtidos pelo gráfico das correlações;
Passo2	Gere os *medoids* iniciais $v_{(0)_k}$, que corresponderá aos padrões de temas que agregarão os termos;

Quadro 2 – Etapas do Método *Fuzzy C-Medoids* (conclusão)

Etapas do Método Fuzzy C-Medoids	
Passo3	Calcule as distâncias euclidianas entre os elementos e o *medoid* do agrupamento: $d(z_j, v_i) = \sqrt{\sum_{j=1}^{p} (z_{ki} - v_{ki})^2}$, onde $v_j = (v_1, v_2, ..., v_k) \in \Re^{cp}$;
Passo4	Gere a função de pertinência pela expressão: $u_{ij} = \exp\left(-\frac{d(z_j, v_i)}{\eta_i}\right)$, onde $\eta_i = \max(d(z_j, v_i))$;
Passo5	Atualize os *medoids*;
Passo6	Calcule a função objetivo $J(V;Z) = \sum_{j=1}^{p}\sum_{i=1}^{k} u_{ij}^m d(z_j, v_i)$. **SE** a função objetivo estiver minimizada; **FIM** **SENÃO** Atualize os *medoids* pela expressão: $v_{0_k} = \frac{\sum_{j=1}^{p} u_{kj}^m z_{kj}}{\sum_{j=1}^{p} u_{kj}^m}$, em função da média ponderada dos valores padronizados correspondentes aos termos, cujas pertinências correspondem aos ponderadores;
Passo7	Volte para o **Passo3**.
Saída: *Agrupamentos segundo temas com respectivos termos inerentes a genômica nutricional.*	

Fonte: A autora, 2019.

4. RESULTADOS E DISCUSSÃO

No processamento foi utilizado o *software RStudio*, que permitiu desenvolver os métodos propostos. É importante verificar o grau de relacionamento dos termos existentes em três documentos pela correlação ordinal de *Spearman* dos escores ranqueados obtidos em função das frequências relativas observadas. A matriz de correlação identificou a interação dos termos e o gráfico de dispersão permitiu visualizar o posicionamento dos termos nos quadrantes do espaço tridimensional. Foi observado que os termos nutrigenoma, genoma, alimento e gene/expressão se destacaram dos demais agrupamentos, mas ficou difícil distinguir demais temas em relação ao restante dos termos. O método *fuzzy C-Means* e o *fuzzy C-Medoids* são opções para discriminar de forma otimizada outros temas. No entanto para viabilizar estes métodos é necessário estabelecer o número provável de agrupamentos que poderiam discriminar os termos dentro de cada conjunto. Neste artigo se optou por seis agrupamentos em função da aglutinação dos termos visualizados na Figura 1. A matriz de correlação subsidiou a análise crítica para verificar se ocorreu relação inversa entre documentos, o que não é esperado na mineração de texto, o que ocorreu entre os Documentos 2 e 3, Tabela 1.

Tabela1 – Determinação do número de grupos pela correlação de *Spearman*

	Documento 1	Documento 2	Documento 3
Documento 1	1,0000	0,1890	0,0362
Documento 2	0,1890	1,0000	-0,0585
Documento 3	0,0362	-0,0585	1,0000

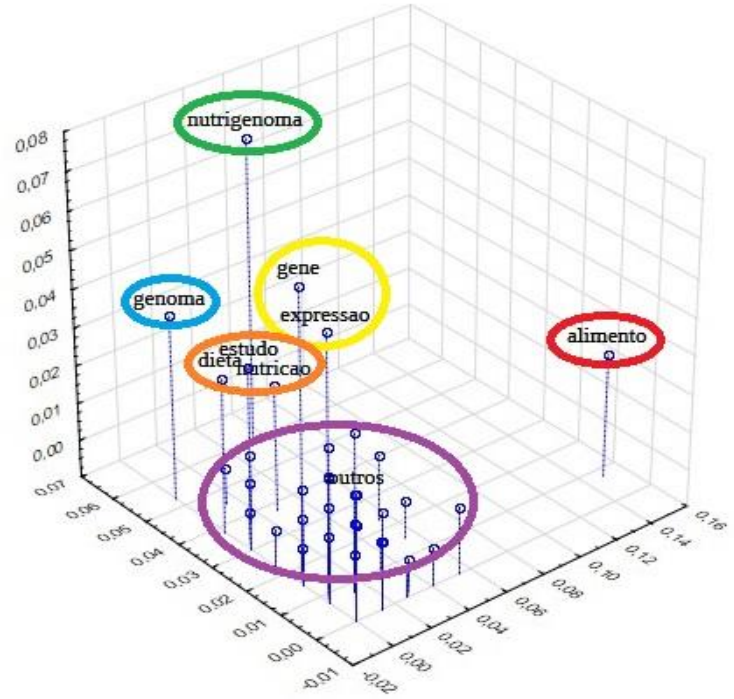

Figura 1 – Determinação do número de grupos pela correlação de *Spearman*

As funções objetivas atingiram os valores de 13,62 e 12,64 para o *fuzzy C-Means* e *C-Medoids,* respectivamente, em 200 iterações e três iterações nesta mesma ordem, mediante a minimização da função objetivo. O processamento do método *C-Means* utilizou a distância euclidiana em relação ao centroide, enquanto no método *C-Medoids*, o Princípio de Extensão *Fuzzy* [29], que adotou um índice relativo gerado pelo quociente de uma função exponencial e o máximo da distância euclidiana observada.

Os resultados obtidos pelos dois métodos foram tratados segundo uma visualização com os agrupamentos e os termos associados. Observou-se que no processamento do *fuzzy C-Medoids* houve a discriminação mais ágil dos termos do que no *fuzzy C-Means*, apontando os termos alimento, nutrigenoma e gene não se agregaram a nenhum dos agrupamentos, embora o termo "gene" poderia ter sido agregado no "nutrigenoma". Estes termos discriminados mostraram alguma ligação aos agrupamentos quando tratados pelo método *Fuzzy C-Means*, entretanto há uma ligação em vizinhanças de grupos.

O Quadro 3 discrimina os termos segundo os algoritmos por agrupamento. Percebe-se que os agrupamentos 4 (*fuzzy C-Means*) e 1 (*fuzzy C-Medoids*), de maior e igual frequência

de termos, dificultam a percepção de cenários. O termo alimento é único no agrupamento 2 do *fuzzy C-Medoids*.

No agrupamento 3 (*fuzzy C-Means*), o algoritmo agregou "alimento" e mais dez de possível relacionamento: ambiente, estresse, exata, fixo, ler, metabolismo, precisa, profissional, remedio, vida. A interpretação deste cenário mostra interação dos termos, porque ambiente pode se referir aos alimentos orgânicos; estresse relacionado ao comportamento alimentar; exata, fixo e precisa está vinculado aos quantitativos dietéticos; ler se relaciona aos *blogs*; metabolismo infere sobre infere sobre anabolismo e catabolismo; profissional e remédio reporta-se aos endocrinologistas, cirurgiões, nutrólogos, nutricionistas e psicólogos; vida é o termo holístico.

Os agrupamentos 1 (*fuzzy C-Means*) e 4 (*fuzzy C-Medoids*) apresentaram sete e seis termos, respectivamente mas apenas três termos coincidentes: ciência, expressão e saúde. O termo ciência, mais abrangente, integra várias áreas do saber: Medicina, Odontologia, Farmácia, Enfermagem, Nutrição, Saúde Coletiva e Educação Física. O termo expressão pode ser percebido como expressão genética, ou seja, processo pelo qual a informação hereditária contida em um gene é processada em um produto genético funcional. Se expressão genética associam-se os termos gene, nutrição, nutrigenoma, vida. Os termos remanescentes precisa e remédio, estão incluídos no agrupamento 3 (*fuzzy C-Means*), o que leva mostrar proximidade entre os dois cenários. O termo entender e gene também são visualizados no agrupamento 6 (*fuzzy C-Medoids*).

Nos agrupamentos 2 (*fuzzy C-Means*) e 6 (*fuzzy C-Medoids*) se percebem termos coincidentes: alterar, baseado, comum, conhecer, determinado, dna, explicar, gerar, importante, ligação, população, principal. O agrupamento 2 tratado pelo *fuzzy C-Means* ficou contido no agrupamento 6 tratado pelo *fuzzy C-Medoids.* Infere-se que houve uma convergência parcial dos algoritmos, embora os termos do agrupamento 6, entender, estresse, exata, fixo, gene, ler, metabolismo e profissional se encaixaram em outros agrupamentos.

O agrupamento 6 do *fuzzy C-Means*, agregou quatro termos: estudo, fator, indivíduo e nutriente. Este processamento não alcançou a construção de um cenário, uma vez que estes termos são díspares e se encontram em outros agrupamentos.

A análise do agrupamento 5 (*fuzzy C-Means*) se afigurou como um subconjunto do agrupamento 3 (*fuzzy C-Medoids*), uma vez que os termos aumenta, bioativo, capacidade, componentes, definição, dieta, efeitos, especifica, genoma, influenciar, interação, metabólico, metaboloma, objetivo, ortomolecular, processo, proteoma, responder, transcrição e transcriptoma são termos análogos para gerar o cenário Nutrigenômica. Os termos remanescentes do agrupamento 3 (*fuzzy C-Medoids*) que foram identificados poderiam conduzir a perda de especificidade do cenário conceitual da Nutrigenômica, tendo em vista a inclusão dos termos: buscar, clínica, conceito, doença, envolver, estudo, fator, genética, gênica, humano, indivíduo, iniciar, interferência, ligação, molecular, nutrição, nutriente, pesquisa, projeto, proteínas e referência.

Quadro 3 – Distribuição dos termos segundo os algoritmos *fuzzy C-Means* e *C-Medoids* (continua)

Agrupamento	Termos agrupados
1 (FCM)	ciência, entender, expressão, gene, nutrição, nutrigenoma, saúde (n = 7)
4 (FCMdd)	ciência, expressão, precisa, remédio, saúde, vida (n = 6)

Quadro 3 – Distribuição dos termos segundo os algoritmos *fuzzy C-Means* e *C-Medoids* (conclusão)

Agrupamento	Termos agrupados
2(FCM)	alterar, baseado, comum, conceito, conhecer, determinado, dna, explicar, gerar, importante, ligação, população, principal (n=13)
6 (FCMdd)	alterar, ambiente, baseado, comum, conhecer, determinado, dna, entender, estresse, exata, explicar, fixo, gene, gerar, importante, ler, metabolismo, população, principal, profissional (n = 19)
3 (FCM)	alimento, ambiente, estresse, exata, fixo, ler, metabolismo, precisa, profissional, remédio, vida (n=11)
2 (FCMdd)	alimento (n = 1)
4 (FCM)	brasil, busca, cardiovasculares, clinica, composto, criado, descobrir, desenvolvimento, doença, envolver, europeu, exame, genética, genica, humano, iniciar, interferência, investigar, laboratório, molecular, mundo, organização, palavra, pesquisa, presente, projeto, proteínas, provocam, realiza, rede, referencia, sangue, sequenciamento, vasos (n = 34)
1 (FCMdd)	brasil, busca, cardiovasculares, clinica, composto, criado, descobrir, desenvolvimento, doença, envolver, europeu, exame, genética, genica, humano, iniciar, interferência, investigar, laboratório, molecular, mundo, organização, palavra, pesquisa, presente, projeto, proteínas, provocam, realiza, rede, referencia, sangue, sequenciamento, vasos (n = 34)
5 (FCM)	aumenta, bioativo, capacidade, componentes, definição, dieta, efeitos, especifica, genoma, influenciar, interação, metabólico, metaboloma, objetivo, ortomolecular, processo, proteoma, responder, rna, transcrição, transcriptoma (n = 21)
3 (FCMdd)	aumenta, bioativo, busca, capacidade, clinica, componentes, conceito, definição, dieta, doença, efeitos, envolver, especifica, estudo, fator, genética, genica, genoma, humano, individuo, influenciar, iniciar, interação, interferência, ligação, metabólico, metaboloma, molecular, nutrição, nutriente, objetivo, ortomolecular, pesquisa, processo, projeto, proteínas, proteoma, referencia, responder, rna, transcrição, transcriptoma (n = 42)
6 (FCM)	estudo, fator, indivíduo, nutriente (n = 4)
-	-

A visualização gráfica dos resultados do *Fuzzy C-Means*, Figura 2, mostrou que o termo alimento não é inerente a Nutrigenômica, se distanciando extremamente dos demais agrupamentos que levaram em consideração termos da Nutrigenômica. O agrupamento 5 se assemelha a uma "árvore semântica" da Nutrigenômica. É importante ressaltar que o agrupamento 1 apresenta termos associados a Nutrigenômica com pouca especificidade. Os agrupamentos 2, 4 e 6 se configuraram difusos sem discriminação dos diferentes termos associados à Nutrigenômica, como apresentado na Figura 1.

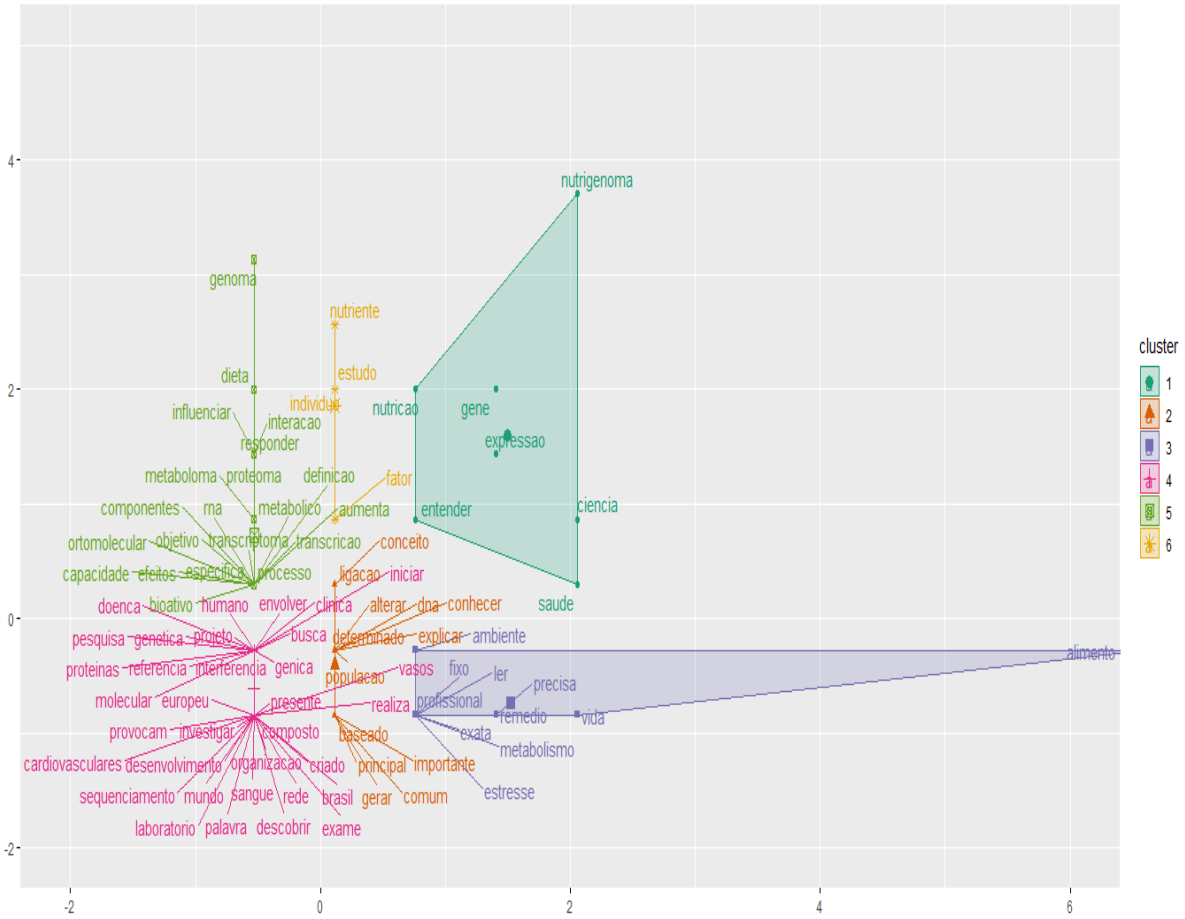

Figura 2 – Gráfico dos termos segundo agrupamento pelo método *fuzzy C-Means*

A interpretação gráfica dos resultados do *Fuzzy C-Medoids*, Figura 3, apesar da rápida convergência em alcançar o mínimo para a função objetivo, não apresentou cenário nítido relativo à Nutrigenômica. Esta figura permitiu visualizar oito agrupamentos ocorrendo duas sobreposições. A interseção dos agrupamentos 1 e 3 mostraram termos superpostos, porém com pouca especificidade: genética, proteína, doença, referência, pesquisa, busca, iniciar, interferência, envolver, projeto, molecular, gênica, clínica e humano. A interseção dos agrupamentos 4 e 5, apenas o termo gene.

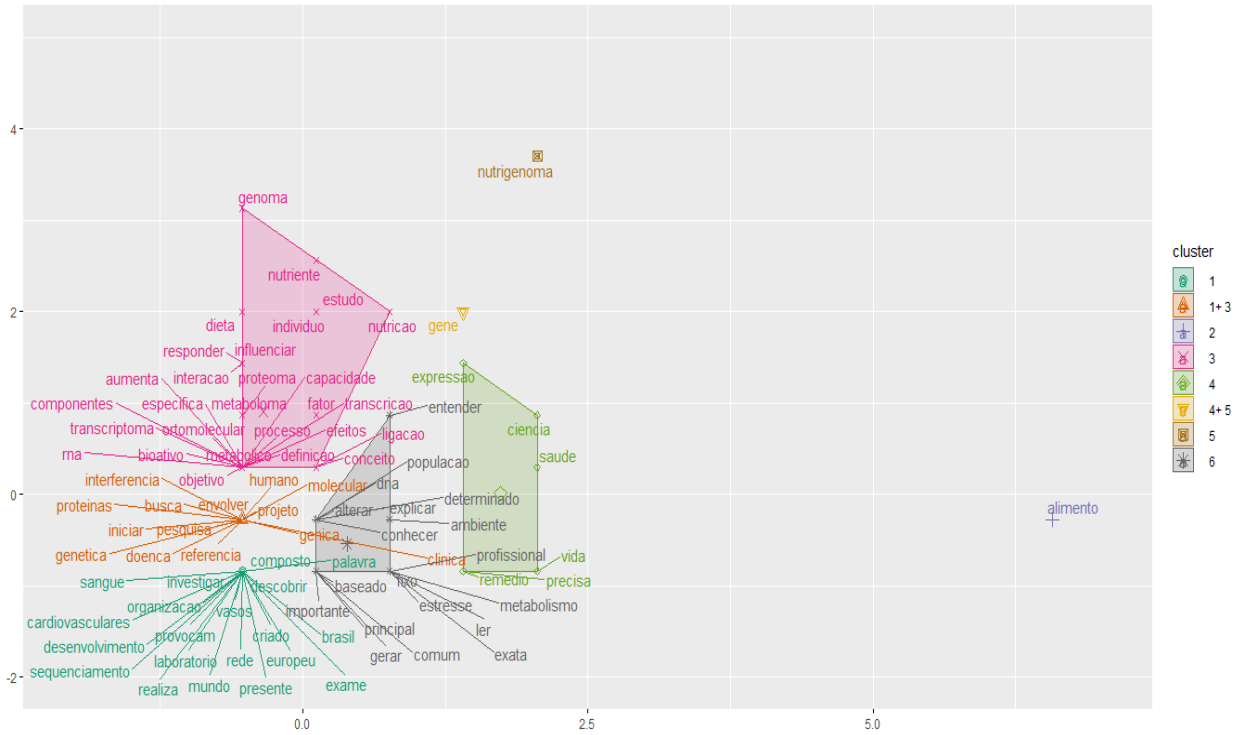

Figura 3 – Gráfico dos termos segundo agrupamento pelo método *fuzzy C-Medoids*

5. CONCLUSÕES

O suporte computacional para a análise de grande volume de dados na busca de informações estruturados sobre um determinado tema se torna viável pelo suporte computacional. Particularmente em Nutrigenômica, cujo acervo é vasto e disperso.

O uso de algoritmos na Mineração de Texto otimiza a busca em função das palavras-chave geradas por eles. No presente estudo o confronto dos algoritmos *fuzzy C-Means* e *fuzzy C-Medoids*, mostrou que o primeiro agregou os termos sob a ótica semântica com maior efetividade.

6. AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

7. REFERÊNCIAS BIBLIOGRÁFICAS

[1] HILBERT, Martin; LÓPEZ, Priscila. The World's Technological Capacity to Store, Communicate and Compute Information. **Sciencexpress Ressearch Article**, Isle of Man, vol. 332, p. 60-65, 2011. Disponível em: http://science.sciencemag.org/content/332/6025/60. Acesso em: 12 fev. 2019.

[2] ROJAS, Alexandre. **Geração Dinâmica de Padrões de Movimentação Rodoviária de Cargas e Veículos por Identificação por Rádio Frequência.** 2013. 180f. Tese (Doutorado) – Curso de Engenharia de Transportes, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013. Disponível em:

http://www.pet.coppe.ufrj.br/images/documentos/teses/2013/Tese_AlexandreRojas.pdf. Acesso em: 13 fev. 2019.

[3] TESSAROLO, Pedro Henrique; MAGALHÃES, Willian Barbosa. **A era do big data no conteúdo digital: os dados estruturados e não estruturados**. Universidade Paranaense, Paranavaí, 2015, 5p. Disponível em: http://web.unipar.br/~seinpar/2015/_include/artigos/Pedro_Henrique_Tessarolo.pdf. Acesso em: 12 fev. 2019.

[4] MAIMON, Oded; ROKACH, Lior. **Data mining and knowledge discovery handbook.** 2 ed. New York: Springer, 2005.

[5] TAN, Ah-Hwee. **Text Mining: The state of the art and the challenges.** Singapore: Nanyang Technological University, 1999. Kent Ridge Digital Labs. Disponível em: http://www3.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf. Acesso em: 30 nov. 2018.

[6] KRASSMANN, Aliane Loureiro *et al.* Analyzing trends in academic papers about ubiquitous virtual worlds in education using text mining. **International Journal for Innovation Education and Research**, Dhaka, p. 167-180, v. 5, abr. 2017. Disponível em: https://ijier.net/index.php/ijier/article/view/650/567. Acesso em: 01 maio 2019.

[7] MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. **Mineração de Textos.** Goiânia: Universidade Federal de Goiás, 2007. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf. Acesso em: 21 mar. 2019.

[8] MENDONÇA, Jonas Henrique; DRUMMOND, Isabela Neves; SANDRI, Sandra Aparecida. Técnicas Baseadas em Redes Neurais Artificiais e Lógica Difusa para Mineração de Textos. *In:* 10° Brazilian Congress on Computacional Intelligence, 10, 2011, Fortaleza, CE. **Anais [...],** Fortaleza: ABRICOM, 2011. Disponível em: http://abricom.org.br/wp-content/uploads/2016/03/st_27.5.pdf. Acesso em: 05 abr. 2019.

[9] MINGOTI, Sueli Aaparecida. **Análise de Dados Através de Métodos de Estatística Multivariada:** Uma Aborgadem Aplicada. 2 ed. Belo Horizonte: UFMG, 2013.

[10] CARRILHO JUNIOR, João Ribeiro. **Desenvolvimento de uma Metodologia para Mineração de Textos**. 2007. 96f. Dissertação (Mestrado) – Curso de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: https://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=11675@1. Acesso em: 07 nov. 2018.

[11] WIVES, Leandro Krug. **Tecnologias de Descoberta de Conhecimento em Textos Aplicadas à Inteligência Competitiva.** 2002. 116f. Tese (Doutorado) – Curso de Computação, Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002. Disponível em: https://seer.ufrgs.br/cadernosdeinformatica/article/view/v1n1p25-28. Acesso em: 10 mar. 2019.

[12] PATIL, Deepa B.; DONGRE, Yashwant V. A fuzzy approach for text mining. **International Journal Mathematical Sciences and Computing**, Hong Kong, p. 34-43, nov. 2015. Disponível em: http://www.mecs-press.org/ijmsc/ijmsc-v1-n4/IJMSC-V1-N4-4.pdf. Acesso em: 13 mar. 2019.

[13] GOULARTE, Fábio Bif. **Método fuzzy para a sumarização automática de texto com base em um modelo extrativo (FSumm).** 2015. 119f. Dissertação (Mestrado) – Curso de Ciência da Computação, Universidade Federal de Santa Catarina, Florianópolis, 2015. Disponível em:

https://repositorio.ufsc.br/handle/123456789/132756. Acesso em: 12 mar. 2019.

[14] DAS, Samarjit. Pattern Recognition using the Fuzzy c-means Technique. **International Journal of Energy, Information and Communications**, North Carolina, v. 4, n. 1, p. 1-14, fev. 2013. Disponível em: https://www.researchgate.net/publication/303150319_Pattern_Recognition_using_the_Fuzzy_c-means_Technique. Acesso em: 11 fev. 2019.

[15] FERREIRA, Rodrigo Nunes. Análise Cluster em Ambiente R: Uma Aplicação dos Algoritmos Fanny, C-means e C-medoids na Classificação dos Municípios Brasileiros Segundo Indicadores de Bem-Estar Social. **Cadernos do LESTE**. Belo Horizonte, dez. 2013. Artigo Científico, p. 151-176. Disponível em: http://igc.ufmg.br/portaldeperiodicos/index.php/leste/article/view/1137/888. Acesso em: 08 mar. 2019.

[16] RODRIGUES, Rodrigo Lins. **Estatística Computacional: Análise de Conglomerados**. Recife, 2018. Universidade Federal Rural de Pernambuco. Disponível em: https://pt.slideshare.net/rodrigomuribec/aula-6-anlise-de-conglomerados?from_action=save. Acesso em: 26 mar. 2019.

[17] HANDL, Julia; KNOWLES, Joshua; KELL, Douglas B. Computational cluster validation in post-genomic data analysis. **Journal Bioinformatics**, Maryland, v. 21, p. 3201-2212, mai. 2005. Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/15914541. Acesso em: 20 abr. 2019.

[18] MALVA, Madalena. **Coeficiente de Correlação Ró de Spearman.** Viseu, 2007. Escola Superior de Tecnologia, Instituto Politécnico de Viseu. Disponível em: http://www.estgv.ipv.pt/PaginasPessoais/malva/TratamentoEstatistico%20de%20dados/Coeficiente%20de%20Correla%C3%A7%C3%A3o%20R%C3%B3%20de%20Spearman.pdf. Acesso em: 17 abr. 2019.

[19] DUNN, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. **Journal of Cybernetics**, p. 32-57, 1973. Disponível em: https://www.tandfonline.com/loi/ucbs19. Acesso em: 28 mar. 2019.

[20] BEZDEK, James. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. 1 ed. New York: Plenum Press, 1981.

[21] GREKOUSIS, George; FOTIS, Yorgos N. A fuzzy for detecting spatiotemporal outliers. **Geoinformatica**, p. 597-619, out. 2011. Disponível em: https://dl.acm.org/citation.cfm?id=2159278. Acesso em: 10 abr. 2019.

[22] PAL, Nikhil R.; BEZDEK, James C. On cluster validity for the fuzzy c-means model. **IEEE Transactions on Fuzzy Systems**, v. 3, p. 370-379, ago. 1995. Disponível em: https://ieeexplore.ieee.org/document/413225. Acesso em: 30 abr. 2019.

[23] YONAMINE, Frank Sussumu *et al.* **Aprendizado não-supervisionado em domínios fuzzy - algoritmo fuzzy c-means.** São Carlos: Universidade Federal de São Carlos, 2002. Disponível em: http://www2.dc.ufscar.br/~carmo/relatorios/RT_Fuzzy_Cmeans_Final.PDF. Acesso em: 11 abr. 2019.

[24] KRISHNAPURAN, Raghu *et. al.* Low-complexity fuzzy relational Clustering algorithms for web mining. **IEEE Transactions on Fuzzy Systems**, v. 9, p. 595-607, ago. 2001. Disponível em: https://ieeexplore.ieee.org/document/940971. Acesso em: 07 abr. 2019.

[25] KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Clustering by Means of Medoids,** 1987. Disponível em: https://www.researchgate.net/publication/243777819_Clustering_by_Means_of_Medoids. Acesso em: 03 abr. 2019.

[26] KAUFMAN, Leonard; ROUSSEEUW, Peter J. **Finding Groups in Data: An Introduction to Cluster Analysis**. New York: John Wiley & Sons, 1990.

[27] KRISHNAPURAN, Raghu; JOSHI, Anupam; YI, Liyu. A Fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. *In:* 1999 IEEE International Fuzzy Systems, 1999, Seoul, South Korea. **Proceedings[...],** Seoul: IEEE Transactions on Fuzzy Systems, 1999. Disponível em: https://ieeexplore.ieee.org/document/790086. Acesso em: 12 abr. 2019.

[28] KRISHNAPURAN, Raghu; KELLER, James M. The Possibilistic C-Means Algorithm: Insights and Recommendations. **IEEE Transactions on Fuzzy Systems**, v. 4, p. 385-393, ago. 1996. Disponível em: https://ieeexplore.ieee.org/document/531779. Acesso em: 02 mai. 2019.

[29] MELO, Gabriel Jesus Alves de. **Princípio de Extensão de Zadeh Aplicado a Funções não Monótonas com dois Parâmetros *Fuzzy***. 2009. 83f. Tese (Doutorado) – Curso de Engenharia de Sistemas (Área de Concentração: Modelagem de Sistemas Biológicos), Universidade Federal de Lavras, Lavras, 2009. Disponível em: http://prpg.ufla.br/_ppg/esistemas//wp-content/uploads/2012/08/3-Dissertacao-Gabriel-J-A-Melo.pdf. Acesso em: 12 mar. 2019.