Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

Influência de Técnicas Não-supervisionadas de Redução de Dimensionalidade para Organização Flexível de Documentos

Beatriz Lima[1], **Fernanda S. Eustáquio**[1], **Tatiane Nogueira**[1]

[1]Instituto de Matemática – Universidade Federal da Bahia (UFBA)
Rua Barão de Jeremoabo, s/n – 40170-115 – Salvador – BA – Brasil

{beatrizlima, fernandase, tatianenogueira}@dcc.ufba.br

Abstract. *Flexible document organization consists of handling uncertainty and imprecision, which are characteristics of natural language's nature and therefore, of texts. In this task, fuzzy clustering has been a powerful allied. However, clustering performance usually is negatively affected by document representation in sparse and high-dimensional vectors, besides the presence of noisy terms. Based on this, the present study seeks to investigate the impact, on fuzzy clustering performance, of dimensionality reduction by using unsupervised methods. The results show that good fuzzy structures are obtained with very few features which can identify the latent semantic aspects within the texts.*

Resumo. *A organização flexível de documentos consiste em agregar tratamento de imprecisão e incerteza, características da natureza da linguagem natural e, por conseguinte, dos textos. Nessa tarefa, o agrupamento fuzzy tem sido um poderoso aliado. Porém, a performance do agrupamento geralmente é afetada negativamente pela representação dos documentos em vetores esparsos e de alta dimensionalidade, além da presença de termos ruidosos. Com base nisso, o presente estudo busca investigar o impacto, na performance do agrupamento fuzzy, da redução de dimensionalidade utilizando técnicas não-supervisionadas. Os resultados mostram que boas estruturas fuzzy são obtidas com muito poucos atributos que conseguem identificar os aspectos semânticos latentes nos textos.*

1. Introdução

A organização e o gerenciamento de documentos digitais tornaram-se tarefas de suma importância nos últimos anos, juntamente com o desenvolvimento de diversos modelos de Sistemas de Recuperação de Informação (SRIs). Dentre eles, está o modelo flexível, no qual os sistemas são capazes de representar e interpretar a subjetividade humana [Bordogna and Pasi 2000].

Documentos textuais são inerentemente incertos e imprecisos, visto que podem ser interpretados de várias formas por diferentes pessoas. Desse modo, para que usuários recuperem a informação contida nesses documentos de maneira mais intuitiva, é necessário que haja uma organização flexível dos mesmos.

Essa flexibilização pode ser obtida por meio de agrupamento fuzzy. Nesse tipo de agrupamento, cada documento pode pertencer a mais de um grupo, com diferentes graus de pertinência, considerando, dessa forma, a possibilidade de existirem características semelhantes entre instâncias de grupos distintos. Além disso, a abordagem fuzzy é dita que consegue tratar as imperfeições características de dados textuais [Kraft et al. 2006].

De modo que o algoritmo de agrupamento possa identificar padrões nos textos, é preciso antes estruturá-los adequadamente. Devido à diversidade de usos, muitas vezes redundantes, das palavras em uma coleção, a representação vetorial para documentos geralmente é esparsa e possui alta dimensionalidade, o que causa impactos negativos no custo computacional e na performance da tarefa de agrupamento. Para contornar essa situação, técnicas de redução de dimensionalidade não-supervisionadas têm sido bastante empregadas. Análise Semântica Latente (*Latent Semantic Analysis* - LSA) [Deerwester et al. 1990, Landauer et al. 1998] e Fatoração de Matriz Não-negativa (*Non-negative Matrix Factorization* - NMF) [Lee and Seung 1999, Lee and Seung 2001] são métodos não-supervisionados de redução comumente aplicados na mineração de textos por conseguirem identificar bem os conceitos semânticos adjacentes nesses dados.

Embora estudos anteriores com LSA e NMF demonstrem que, de maneira geral, os resultados melhoram à medida que a quantidade de dimensões aumenta, até alcançarem um resultado ótimo geralmente em torno de algumas centenas de dimensões [Deerwester et al. 1990, Schütze and Silverstein 1997, Tsuge et al. 2001], nossos experimentos com agrupamento fuzzy mostram um comportamento contrário. Além disso, não existem muitos casos na literatura para análise da influência, em agrupamento fuzzy, de técnicas mais robustas no pré-processamento de documentos, como investigado por este trabalho, apesar de ser de grande relevância para a construção de SRIs flexíveis.

Visando apresentar a investigação realizada, este artigo apresenta a seguinte estrutura. Na seção 2, é fornecida uma visão geral das técnicas LSA e NMF. A Seção 3 descreve brevemente o algoritmo de agrupamento fuzzy mais conhecido e escolhido para realizar as investigações aqui apresentadas, o Fuzzy C-Means (FCM), bem como os índices utilizados para avaliar os resultados. Os experimentos e seus resultados são discutidos na Seção 4. Por fim, na Seção 5, são feitas as considerações finais.

2. Redução de Dimensionalidade Não-supervisionada

A "maldição" da dimensionalidade é um dos maiores desafios associados à descoberta de conhecimento em textos [Zervas and Ruger 1999]. A representação de documentos em vetores de alta dimensionalidade torna o agrupamento mais difícil de ser realizado visto que quaisquer pares de vetores desse tipo tendem a apresentar distâncias constantes uns dos outros no espaço vetorial. Motivadas por esse problema, várias técnicas têm sido investigadas com o intuito de reduzir o número de dimensões no modelo espaço vetorial.

Dentre os métodos de redução de dimensionalidade, estão os não-supervisionados, que derivam novos atributos, em menor quantidade, a partir de um vetor inicial de atributos, por meio de relações adjacentes observadas do comportamento dos dados. Essa abordagem é dita não-supervisionada visto que nenhuma informação acerca de rótulos dos dados é utilizada.

Análise Semântica Latente (LSA) e *Fatoração de Matriz Não-negativa* (NMF) são dois exemplos bem conhecidos de técnicas não-supervisionadas que têm sido aplicadas com sucesso em análises textuais, como, por exemplo, na sumarização automática de textos [Lee et al. 2009], recuperação de informação [Deerwester et al. 1990, Tsuge et al. 2001, Muflikhah and Baharudin 2009] e agrupamento de documentos [Schütze and Silverstein 1997, Shafiei et al. 2007, Yang and Watada 2011].

Ambos os métodos consistem em aglomerar termos semanticamente similares em

um mesmo conceito latente, permitindo, por exemplo, associar a documentos atributos que não estavam antes associados devido à variação de usos de uma mesma palavra através do emprego de sinônimos e polissemia. Além disso, LSA e NMF conseguem reduzir a influência de termos ruidosos. Portanto, mais do que diminuir o custo computacional por causa do uso de uma menor quantidade de termos, essas técnicas auxiliam a obter um conjunto de atributos melhores.

2.1. Análise Semântica Latente (LSA)

Suponhamos que um determinado corpus seja composto por t atributos e d documentos e seja representado por uma matriz termos-documentos $X \in \mathbb{R}^{t \times d}$. LSA aplica a Decomposição em Valores Singulares (*Singular Values Decomposition* - SVD) tal que $X = T_{t \times p} S_{p \times p} D^t_{p \times d}$, onde $p = \min(t, d)$, T e D são matrizes ortogonais e S é uma matriz diagonal de *valores singulares* positivos e ordenados em decrescência.

A redução de dimensionalidade ocorre por meio de uma aproximação de baixo posto (*rank*) tal que $X \approx \hat{X} = T_{t \times k} S_{k \times k} D^t_{k \times d}$. Nesse caso, considera-se que apenas os k maiores valores singulares em S, sendo $k \ll p$, são suficientes para conseguir uma boa aproximação de X. Essa abordagem conhecida como ***SVD truncada*** permite que X seja transformada em uma nova matriz termos-documentos $\hat{X}$ de *rank* k. Assim, cada documento passa a ser descrito como uma combinação linear dos k componentes LSA.

2.2. Fatoração de Matriz Não-negativa (NMF)

Enquanto que no LSA não existem restrições para os valores nas matrizes T e D, o método NMF gera apenas matrizes não-negativas no processo de decomposição. Por esse motivo, o NMF pode ser considerado mais intuitivo, principalmente para áreas em que essa restrição é importante para interpretação dos novos atributos, como em análise de imagens e mineração de textos [Lee and Seung 1999].

Dada uma matriz $X_{t \times d}$ não-negativa, NMF obtém a decomposição aproximada $X \approx W_{t \times k} H_{k \times d}$ tal que $W, H \in \mathbb{R}_+$. Como os vetores em W não são ortogonais, pode haver sobreposição entre os novos atributos extraídos, também chamados de tópicos ou conceitos latentes.

3. Agrupamento fuzzy

Para organizar as coleções de documentos de maneira flexível neste trabalho, foi escolhido o algoritmo mais utilizado de agrupamento fuzzy, Fuzzy C-Means (FCM) [Bezdek 1981]. O FCM determina a melhor partição fuzzy ao minimizar sua função objetivo, onde os objetos, aqui tratados como documentos, são atribuídos aos grupos (clusters) através do seu grau de pertinência em cada um dos c clusters. A soma dos graus de pertinência de todos os documentos em um cluster é igual a 1, assim como a soma dos graus de pertinência de um documento em todos os clusters. A dissimilaridade entre um documento e um protótipo foi medida, neste trabalho, utilizando a distância Euclidiana.

No presente trabalho, o FCM foi executado para cada conjunto de dados utilizando como parâmetros o valor de c igual ao número de classes com o qual cada corpus foi rotulado previamente; número de inicializações randômicas $RS = 10$ para limitar o risco de acertar um ótimo local; critério de convergência $conv = 0,01$ e valores padrão dos parâmetros fator de fuzzificação $m = 2$ e número máximo de iterações $maxit = 10^6$ adotados pela função ***FKM*** do pacote *fclust* do R [Ferraro and Giordani 2015].

3.1. Índices de Validação

Para avaliar as partições fuzzy obtidas pelo FCM nos experimentos, foram utilizados os índices de validação de agrupamento fuzzy mais comumente empregados: Coeficiente da Partição (PC) [Bezdek 1974b], Entropia da Partição (PE) [Bezdek 1974a], Coeficiente da Partição Modificado (MPC) [Dave 1996], Xie-Beni (XB) [Xie and Beni 1991] e Silhueta Fuzzy (SF) [Campello and Hruschka 2006]. Cada um deles é explicado brevemente a seguir.

PC é um índice de maximização e pode assumir valores entre $[1/c, 1]$.

$$PC = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}(A_i(\mathbf{d}_j))^2 \tag{1}$$

PE é um índice de minimização que mede o montante de fuzzificação em uma partição U e pode assumir valores entre $[0, \log_a c]$ onde, neste trabalho, foi utilizado o valor de $a = e$.

$$PE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}A_i(\mathbf{d}_j)\log_a(A_i(\mathbf{d}_j)) \tag{2}$$

MPC é um índice de maximização e foi proposto para corrigir a tendência monotônica do PC. Os resultados obtidos com o uso desse índice variam entre $[0, 1]$.

$$MPC = 1 - \frac{c}{c-1}(1 - PC) \tag{3}$$

XB é um índice de minimização onde um valor de XB pequeno indica que os clusters são compactos e bem separados.

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}(A_i(\mathbf{d}_j))^m\|\mathbf{d}_j - \mathbf{v}_i\|^2}{n \times min_{j\neq i}\|\mathbf{v}_i - \mathbf{v}_j\|^2} \tag{4}$$

SF é a versão fuzzy do índice Silhueta. É um índice de maximização que considera os dois clusters em que d_j tem os dois maiores graus de pertinência.

$$S(d_i) = \frac{\beta(\mathbf{d}_j, g_i) - \delta(\mathbf{d}_j, g_i)}{max\left\{\delta(\mathbf{d}_j, g_i), \beta(\mathbf{d}_j, g_i)\right\}} \tag{5}$$

$$SF = \frac{\sum_{j=1}^{n}(A_1(\mathbf{d}_j) - A_2(\mathbf{d}_j))S(\mathbf{d}_j)}{\sum_{j=1}^{n}(A_1(\mathbf{d}_j) - A_2(\mathbf{d}_j))} \tag{6}$$

onde $\mathbf{d}_j$ pertence ao cluster g_i, $g_i \in (g_1, g_2, ..., g_c)$. $\delta(\mathbf{d}_j, g_i)$ é a média da distância entre $\mathbf{d}_j$ e todos os documentos pertencentes a g_i, i.e. a distância intra-cluster. $\beta(\mathbf{d}_j, g_i)$ é a distância entre $\mathbf{d}_j$ e seu vizinho mais próximo a g_i, i.e. a distância inter-cluster.

4. Experimentos

Os experimentos[1] foram conduzidos com 4 bases reais e bem conhecidas (Tabela 1)[Rossi et al. 2013], representadas com variadas dimensões. A nomenclatura utilizada para essas representações seguem o padrão LSA-k e NMF-k, em que k corresponde ao *rank* das matrizes reduzidas, isto é, k é a quantidade de atributos extraídos pelas técnicas de redução. Empiricamente alguns valores para k foram testados e, de maneira geral, melhores resultados foram obtidos com valores muito baixos. Dessa forma, foi definido que k deveria variar entre 2 e 10. Ou seja, a representação LSA-10, por exemplo, corresponde a uma matriz documentos-termos com 10 conceitos latentes extraídos pela técnica LSA. Assim sendo, os resultados obtidos foram analisados sob a seguinte perspectiva:

É possível obter uma estrutura satisfatória de grupos fuzzy utilizando muito poucos atributos extraídos pelos métodos LSA e NMF?

Acreditamos que o questionamento acima leva a realizar escolhas importantes para se obter uma organização e recuperação flexível dos documentos bem sucedidas.

4.1. Corpora

As principais características das quatro bases[2] escolhidas para realizar os experimentos estão descritas na Tabela 1.

Tabela 1. Características das coleções

Base	Domínio	# documentos	# atributos	# classes
CSTR	Científico	299	1725	4
IrishSentiment	Análise de Sentimentos	1660	8658	3
Hitech	Notícias	2301	12941	6
La1s	Notícias	3204	13195	6

Apesar das coleções estarem rotuladas, essa informação só é utilizada na definição do número de grupos que é passado para o algoritmo FCM. Além disso, é da natureza dos textos serem incertos e imprecisos, isto é, um mesmo texto pode discorrer sobre vários temas com diferentes graus de abordagens. Na base CSTR, por exemplo, um relatório técnico pode ter um foco maior na área de Robótica, sendo por isso rotulado com a classe *Robotics*, mas tratar eventualmente de conceitos de Sistemas (classe *Systems*). É partindo dessa intuição que acreditamos que o agrupamento fuzzy é adequado para esse cenário.

Antes de aplicar os métodos LSA e NMF, os documentos foram convertidos em vetores de atributos unigramas, os quais foram extraídos após remoção de *stopwords* e *stemming*. Por fim, o esquema de peso escolhido foi o *tf-idf*, devido a sua capacidade de reduzir a importância de termos que são muito comuns na coleção e também devido aos bons resultados alcançados com esse esquema no agrupamento de textos [Singh et al. 2011].

[1]Foram utilizados nos experimentos os pacotes do R *lsa* [Wild 2015], *NMF* [Gaujoux and Seoighe 2010] e *fclust* [Ferraro and Giordani 2015].

[2]As bases estão disponíveis no repositório de coleções textuais do LABIC-USP em `http://sites.labic.icmc.usp.br/text_collections/`.

Esses vetores iniciais formam a representação *baseline* denominada TFIDF, que aqui é tratada como a matriz documentos-termos sem redução de dimensionalidade. A matriz TFIDF de cada coleção possui a quantidade de atributos descrita na Tabela 1. Sendo assim, para cada base foram realizados testes com 19 representações diferentes: LSA-2, ..., LSA-10, NMF-2, ..., NMF-10 e TFIDF.

4.2. Resultados

Os índices de validação de agrupamento fuzzy descritos na Seção 3.1 foram utilizados para avaliar os agrupamentos realizados sobre cada uma das bases com as 19 diferentes representações explicadas anteriormente. Os resultados obtidos podem ser visualizados nos gráficos da Figura 1.

É possível verificar pela Figura 1 que os índices de validação PC, PE e MPC não conseguiram identificar boas estruturas de clusters em nenhuma coleção agrupada com a representação TFIDF, dado que todos estes índices obtiveram valores muito próximos dos seus respectivos limites. Para os índices de maximização PC e MPC, os valores encontrados foram muito próximos dos respectivos limites inferiores de $1/c$ (CSTR = $1/4$, IrishSentiment = $1/3$, Hitech e La1s = $1/6$) e 0. Para o índice de minimização PE, os valores encontrados para todas as coleções também foram muito próximos do limite superior de $\ln c$ (CSTR = $\ln 4$, IrishSentiment = $\ln 3$, Hitech e La1s = $\ln 6$).

Os resultados dos índices, ao avaliarem coleções representadas pelo TFIDF, foram inferiores aos gerados pelos métodos LSA e NMF, com exceção aos agrupamentos das coleções CSTR e IrishSentiment avaliadas pelo SF e La1s avaliada pelo XB, como pode ser visto na Tabela 2. Isso confirma a capacidade de ambos LSA e NMF em descobrirem conceitos semânticos intrínsecos nos dados que descrevem melhor as características dos mesmos do que os atributos iniciais TFIDF, inclusive no contexto de organização flexível desses documentos.

Tabela 2. Representações avaliadas com os melhores e piores resultados

	PC		PE		MPC		SF		XB	
	Melhor	Pior	Melhor	Pior	Melhor	Pior	Melhor	Pior	Melhor	Pior
CSTR	LSA-3	TFIDF	LSA-3	TFIDF	LSA-3	TFIDF	NMF-4	NMF-8	NMF-2	TFIDF
IrishSentiment	LSA-2	TFIDF	LSA-2	TFIDF	LSA-2	TFIDF	LSA-8	NMF-9	NMF-2	TFIDF
Hitech	LSA-2	TFIDF	LSA-2	TFIDF	LSA-2	TFIDF	NMF-2	LSA-9	NMF-2	TFIDF
La1s	LSA-2	TFIDF	LSA-2	TFIDF	LSA-2	TFIDF	NMF-2	LSA-8	NMF-2	NMF-9

Contudo, não foi somente com a avaliação do agrupamento nas coleções representadas pelo TFIDF que os índices não obtiveram valores satisfatórios. Pelos valores apresentados pelos índices PC, PE e MPC (Figura 1) percebe-se que, a partir de um número k de dimensões, as representações com LSA e com NMF tiveram comportamento semelhante ao TFIDF ao apresentarem novamente valores muito próximos aos limites destes índices para todas as coleções de documentos, com exceção da La1s que apresentou valores satisfatórios a partir de $k = 10$ para ambos LSA e NMF.

A partir da Figura 1, foi possível identificar os valores de k em que estes começam a encontrar uma boa estrutura nas coleções. Esta mudança pode ser vista a partir do ponto em que os valores dos índices PC, PE, MPC começam a se distanciar no mínimo em 0,01

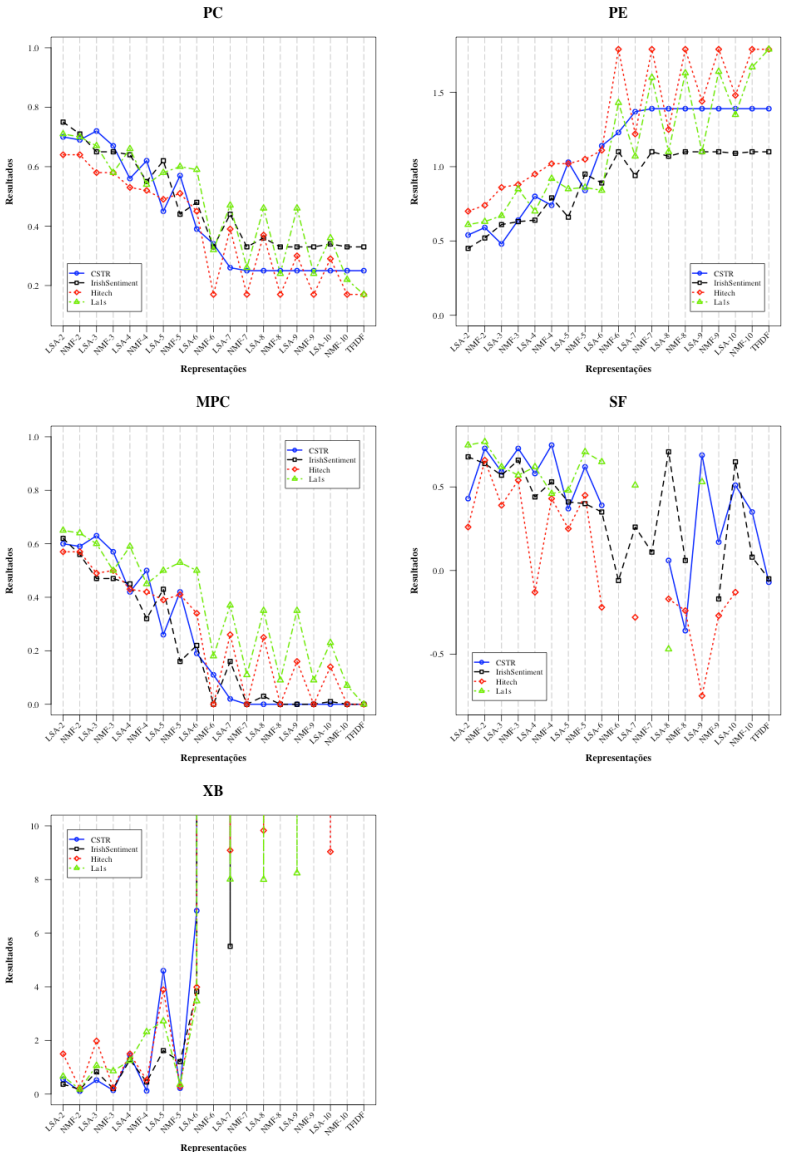

Figura 1. Cada gráfico corresponde aos resultados obtidos por um índice de validação. As linhas descontínuas entre as representações no gráfico do SF indicam valores não numéricos (NaN). Devido a alta amplitude dos valores obtidos pelo XB (com valor mínimo obtido de $0,11$ para a coleção CSTR e valor máximo $4,63\text{e}{+13}$ para o La1s), seu gráfico foi limitado de 0 a 10 para que o comportamento apresentado pelas representações para cada coleção fosse perceptível. As bases estão representadas com cores e linhas diferentes, como mostram as legendas. No eixo x se encontram os nomes de cada uma das 19 representações em ordem crescente no número de dimensões.

dos seus limites. Para o XB, os valores de k foram identificados a partir da diferença exponencial entre as representações com k e $k-1$ dimensões. Os respectivos valores de k para cada coleção dado os índices PC, PE, MPC e XB são apresentados na Tabela 3.

Tabela 3. Número k de dimensões para o qual os valores dos índices apresentaram uma melhora significativa. O índice SF não foi considerado por não ter apresentado valores discrepantes ao variar o valor de k.

	PC		PE		MPC		XB	
	LSA	NMF	LSA	NMF	LSA	NMF	LSA	NMF
CSTR	$k < 8$	$k < 7$	$k < 8$	$k < 7$	$k < 8$	$k < 7$	$k < 7$	$k < 6$
IrishSentiment	$k < 9$	$k < 6$	$k < 9$	$k < 6$	$k < 9$	$k < 6$	$k < 8$	$k < 6$
Hitech	$k \leq 10$	$k < 6$	$k \leq 10$	$k < 6$	$k \leq 10$	$k < 6$	$k \leq 10$	$k < 6$
La1s	$k \leq 10$	$k \leq 10$	$k \leq 10$	$k \leq 10$	$k \leq 10$	$k \leq 10$	$k < 10$	$k < 6$

Pela Tabela 3 pode-se assumir que para as coleções representadas pelo NMF, o valor de $k = 5$ já permite um bom agrupamento assim como o valor de $k = 6$ para as coleções representadas pelo LSA.

Quando comparados os números de dimensões, principalmente com a avaliação pelos índices PC, PE e MPC que foram unânimes ao avaliarem as mesmas representações como as de resultado superior e inferior, percebe-se que os resultados pioram à medida que a quantidade de dimensões cresce. A maior parte das melhores estruturas fuzzy foram encontrados em um espaço vetorial com apenas $k = 2$ dimensões, como mostra a Tabela 2.

5. Conclusão

Organizar documentos de maneira flexível é uma alternativa importante para uma recuperação da informação que atenda melhor às necessidades dos usuários. Os resultados discutidos na Seção 4.2 são encorajadores por mostrarem que é possível obter uma organização flexível para coleções de documentos utilizando pouquíssimos atributos. Isso traz melhorias consideráveis no tempo de processamento dos documentos e, por conseguinte, na performance de um SRI flexível.

Além disso, pôde-se atestar a superioridade dos conceitos latentes obtidos com as técnicas LSA e NMF, provavelmente devido à capacidade de lidarem com sinônimos e termos polissêmicos. Desse modo, os documentos textuais podem ser representados de maneira mais realística, sobressaindo as suas características naturais de imprecisão e incerteza, o que torna a organização flexível com agrupamento fuzzy bastante adequada nesse contexto.

Para o futuro, é importante tornar o contexto completamente não-supervisionado e testar diversas quantidades de grupos em vez de usar a informação do número de classes. Nesse sentido é promissor investigar a extração de descritores dos grupos fuzzy a fim de obter uma análise mais detalhada dos mesmos.

Agradecimentos

As autoras agradecem ao suporte financeiro concedido pelo CNPq e pela FAPESB.

Referências

Bezdek, J. C. (1974a). Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3(3):58–73.

Bezdek, J. C. (1974b). Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, 1(1):57–71.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Bordogna, G. and Pasi, G. (2000). Modeling vagueness in information retrieval. In *Lectures on information retrieval*, pages 207–241. Springer.

Campello, R. and Hruschka, E. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, 157(21):2858 – 2875.

Dave, R. N. (1996). Validating fuzzy partitions obtained through c-shells clustering. *Pattern Recognition Letter*, 17(6):613–623.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Ferraro, M. and Giordani, P. (2015). A toolbox for fuzzy clustering using the r programming language. *Fuzzy Sets and Systems*, 279:1–16.

Gaujoux, R. and Seoighe, C. (2010). A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1):367.

Kraft, D. H., Pasi, G., and Bordogna, G. (2006). Vagueness and uncertainty in information retrieval: how can fuzzy sets help? In *Proceedings of the 2006 international workshop on Research issues in digital libraries*, page 3. ACM.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562.

Lee, J.-H., Park, S., Ahn, C.-M., and Kim, D. (2009). Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, 45(1):20–34.

Muflikhah, L. and Baharudin, B. (2009). High performance in minimizing of term-document matrix representation for document clustering. In *Innovative Technologies in Intelligent Systems and Industrial Applications, 2009. CITISIA 2009*, pages 225–229. IEEE.

Rossi, R. G., Marcacini, R. M., and Rezende, S. O. (2013). Benchmarking text collections for classification and clustering tasks. Technical report, Institute of Mathematics and Computer Sciences, University of Sao Paulo.

Schütze, H. and Silverstein, C. (1997). Projections for efficient document clustering. In *ACM SIGIR Forum*, volume 31, pages 74–81. ACM.

Shafiei, M., Wang, S., Zhang, R., Milios, E., Tang, B., Tougas, J., and Spiteri, R. (2007). Document representation and dimension reduction for text clustering. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pages 770–779. IEEE.

Singh, V. K., Tiwari, N., and Garg, S. (2011). Document clustering using k-means, heuristic k-means and fuzzy c-means. In *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*, pages 297–301. IEEE.

Tsuge, S., Shishibori, M., Kuroiwa, S., and Kita, K. (2001). Dimensionality reduction using non-negative matrix factorization for information retrieval. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 2, pages 960–965. IEEE.

Wild, F. (2015). *lsa: Latent Semantic Analysis*. R package version 0.73.1.

Xie, X. L. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847.

Yang, J. and Watada, J. (2011). Decomposition of term-document matrix representation for clustering analysis. In *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, pages 976–983. IEEE.

Zervas, G. and Ruger, S. M. (1999). The curse of dimensionality and document clustering. In *Microengineering in Optics and Optoelectronics (Ref. No. 1999/187), IEE Colloquium on*, pages 19–19. IET.