Identificação de Temas em Redes Sociais por meio de técnicas de agrupamento

Marjori N. M. Klinczak[1], **Celso A. Kaestner**[2]

[1]Mosaic Web

[2]Universidade Tecnologica Federal do Paraná (UTFPR)

{contato@mosaicweb.com.br, celsokaestner@utfpr.edu.br

Abstract. *Recent years have been marked by the emergence of various social media, from Orkut to Facebook, including Twitter, Youtube, Google+ and many others: each offers new features to attract more users. These social media generate a large amount of data which if processed properly can be used to identify trends, patterns and changes. The objective of this work is the discovery of the key topics in a social network discussion, characterized as groups of relevant terms restricted to a context, and the study of its evolution over time. To do so we use procedures based on Data Mining and Text Processing. At first, text processing techniques are used to identify the most relevant terms that appear in the text messages of the social network. Then these terms are clustered using the classical k-means and k-medoids algorithms, and also the recent NMF (Non-negative Matrix Factorization) algorithm. Finally, we associate the most relevant terms of theses clusters to characterize the main themes of the considered messages. The proposal was evaluated on the Twitter network, using datasets of tweets of several initial contexts. The results show the feasibility of the proposal in order to identify the relevant topics of this social network in the desire initial context.*

Resumo. *Os anos recentes foram marcados pelo surgimento de várias mídias sociais, do Orkut ao Facebook, incluindo Twitter, Youtube, Google+ e muitos outros: cada um oferece novos recursos para atrair mais usuários. Estas mídias sociais geram uma grande quantidade de dados que se processada corretamente pode ser utilizada para identificar tendências, padrões e mudanças. O objetivo deste trabalho é a descoberta de temas-chave em uma discussão de redes sociais, caracterizada como grupos de termos relevantes restritos a um contexto, e o estudo de sua evolução ao longo do tempo. Para isso, utilizamos procedimentos baseados em mineração de dados e processamento de texto. No início, as técnicas de processamento de texto são usadas para identificar os termos mais relevantes que aparecem nas mensagens de texto da rede social. Em seguida, esses termos são agrupados usando os algoritmos k-means e k-medoids clássicos, e também o recente algoritmo NMF (Non-negative Matrix Factorization). Finalmente, associamos os termos mais relevantes dos agrupamentos de documentos para caracterizar os principais temas das mensagens consideradas. A proposta foi avaliada na rede Twitter, usando conjuntos de dados de tweets de vários contextos iniciais. Os resultados mostram a viabilidade da proposta, a fim de identificar os tópicos relevantes desta rede social no contexto inicial fornecido.*

1. Introdução

Segundo [Ellison 2007] uma rede social online é uma plataforma que oferece um espaço de comunicação e de interação digital a conjuntos de pessoas com necessidades e interesses semelhantes. Redes sociais, tais como Twitter, Facebook, Myspace e Google+, entre outras, estão mudando o comportamento humano [Nel 2011] e incentivando o usuário a expressar seus pensamentos, sentimentos, opiniões e partilhar detalhes de sua vida em curtas mensagens de texto que podem ser enviadas por telefones celulares ou via web [Paul 2011], [Naaman 2010], [Lerman 2010], [Asur 2010] e [Java 2007].

Além disto, pequenas mensagens de texto são prevalentes em várias aplicações web, como microblogs e mensagens instantâneas [Xiaohui 2013], e essas mensagens pouco a pouco estão transformando o modo pelo qual nos comunicamos e vêm desempenhando um papel importante em nossas vidas sociais [Hu 2012] e [Java 2007], pois serviços de rede social como Facebook, Myspace e Twitter são hoje consideradas ferramentas importantes de comunicação para muitos usuários online [Liangjie 2011].

A alta conectividade e a resposta muitas vezes instantânea alimenta a distribuição e disseminação da informação entre os usuários das redes sociais [Gupta 2012], cujo número continua crescendo exponencialmente [Liu 2012], o que oferece uma oportunidade única para analisar a propagação da informação [Lerman 2010]. O sucesso das redes sociais online gerou um problema de difusão de informação em larga escala [Kwak], pois a quantidade de informações compartilhadas pode deixar inviável a procura por determinado tópico. Outro problema encontrado é a presença de ruídos e de inconsistência na escrita, pois devido ao pequeno espaço disponível para escrever a mensagem, os usuários utilizam amplamente *emoticons* e abreviaturas para se expressarem.

Esse trabalho tem por objetivo identificar os temas mais discutidos em um recorte das mensagens do Twitter dentro de um contexto definido pelo usuário, relacionando os termos associados a esses temas. Para tanto será utilizado o algoritmo de agrupamento NMF (*Non negative Matrix Factorization*) [Lee and Seung 2000] – cujos resultados serão comparados com os obtidos do uso das técnicas tradicionais de agrupamento k-means e k-medoids.

No contexto desse trabalho entende-se por tema um assunto que está sendo discutido nas mensagens do Twitter, definido como um conjunto de termos (ou palavras) importantes identificados nas mensagens e caracterizado e visualizado por um conjunto de vetores de termos ou por uma nuvem de palavras, obtidos a partir de um contexto inicial dado.

A técnica de agrupamento NMF tem sido empregada em vários trabalhos e existe desde a década de 90 [Lee and Seung 2000], porém não foi encontrada na literatura sua aplicação às mensagens de redes sociais antes da pesquisa dos autores em [Klinczak 2015] e [Klinczak 2016]. Várias opções de ponderação de termos advindas da área de Recuperação de Informações (booleana, TF, TF-IDF) são empregadas no pré-processamento textual, bem como várias medidas de distância entre os vetores de termos (distância euclidiana e medida do cosseno). A comparação com os resultados obtidos com os algoritmos tradicionais de agrupamento empregam métricas intra e inter agrupamento (WSS e BSS), como detalhado adiante.

2. Trabalhos Relacionados

Pesquisas de diversas áreas analisaram o conteúdo de redes sociais de forma a extrair conhecimento para seus domínios; nesta seção são apresentados alguns destes trabalhos correlacionados.

Milhares de pessoas ficaram sabendo da morte de Osama Bin Laden minutos antes da comunicação oficial devido ao Twitter; os autores [Hu 2012] buscaram entender o papel desempenhado pelo Twitter no episódio. Para isso extraíram *tweets* sobre Osama Bin Laden postados na janela de 2 horas antes e depois das notícias oficiais, e também analisaram os links compartilhados pelo Twitter de acordo com seu conteúdo. Os *tweets* em inglês foram priorizados, coletando-se 614.976 mensagens contendo o termo "*laden*" durante quase 2 horas; as mensagens foram classificadas em correto, incorreto e irrelevante, utilizando o classificador SVM. Os autores também buscaram descobrir de onde partiu a primeira mensagem sobre o assunto, para isso examinaram os *tweets* entre 10:20pm e 10:45pm, concluindo que havia sido Keith Urban, ex-secretário de defesa dos EUA o primeiro a fazer o anúncio nesta rede social. Desta forma concluiu-se que a maior parte da informação consumida é criada por um pequeno grupo de usuários, conhecidos como usuários de "elite", e que as demais pessoas tendem a acreditar que uma informação é verdadeira quando ela é apresentada por especialistas no assunto ou celebridades.

A ideia chave do trabalho proposto por [Xiaohui 2013] foi a de aprender os tópicos principais de mensagens explorando a correlação dos termos nos dados. Primeiramente foi computada a correlação dos termos em pequenas mensagens, onde cada termo foi representado por uma lista de termos relacionados, através da construção de uma matriz de co-ocorrência. Considerou-se que os dados resumem todo tipo de informação, desde informações pessoais até novos eventos, buscando-se descobrir tópicos emergentes em mídias sociais com o objetivo de personalizar recomendações. A identificação de tópicos foi feita com a aplicação do algoritmo NMF sobre a matriz de co-ocorrência dos termos. Os testes foram aplicados em 3 bases de dados, mostrando que o método aplicado provê resultados substancialmente superiores aos da *baseline*.

Outro trabalho similar é o de [Klinczak 2015] a respeito do escândalo de corrupção envolvendo a FIFA, que diferentemente de outros trabalhos, compara diretamente a performance de alguns algoritmos de agrupamento (k-means, k-medoids e NMF) com a ponderação de termos TF-IDF, com alguns dados obtidos do Twitter utilizando as *hashtags* "*fifa*" e "*fifagate*" como contexto inicial. Depois de aplicadas as técnicas de pré-processamento restaram 2.460 *tweets* e a aplicação dos algoritmos apresentaram resultados similares na maior parte dos casos, sendo que o algoritmo NMF apresentou o melhor resultado devido que o mesmo permite que um termo apareça em mais de um agrupamento com diferentes pesos.

3. Metodologia

3.1. Etapas de extração e tratamento da informação

Para usar convenientemente as mensagens do Twitter para a extração de informações é proposto um método composto de uma série de etapas. Inicialmente, os *tweets* são gravados utilizando-se uma API específica, disponibilizada pela própria plataforma, que inclui diversos filtros, como a presença de *hashtags*, o idioma empregado nas mensagens,

restrições de localização geográfica, informações sobre *retweets*, entre outros. Além da mensagem de texto em si, o registro obtido inclui meta-atributos como a identificação ("id") do usuário, a hora e data da postagem do *tweet*, a localização geográfica do emissor (quando disponível), e alguns metadados como imagens e links, que podem ser utilizados para propósitos específicos. A extração das mensagens foi feita com auxílio da biblioteca twitteR da linguagem R, sendo armazenados apenas os *tweets* associados às *hashtags* que definem o contexto inicial.

Na segunda etapa a base de *tweets* armazenada sofreu um pré-processamento textual de acordo com os seguintes passos: (a) identificação das unidades textuais, que nesse caso resumem-se ao conteúdo textual das mensagens; (b) *case-folding*: padronização das ocorrências de caracteres e eventuais conversões de codificação; (c) remoção de *stopwords*, elementos textuais muito frequentes que praticamente não carregam nenhuma informação semântica e são eliminados; uma lista típica dessas palavras inclui artigos, preposições, conjunções e, em algumas aplicações, números; e (d) *stemming*: procedimento que tem por objetivo obter os radicais dos elementos textuais, de forma a conectar elementos de semântica similar: sufixos e prefixos são eliminados, plurais e variações verbais do mesmo elemento são reduzidas para uma forma única. Para *stemming* utilizou-se o já bem conhecido algoritmo de [Porter 1980], e foram considerados apenas termos com mais de dois caracteres. Como resultado desta etapa de pré-processamento, cada *tweet* é agora representado por uma série de radicais de elementos textuais, usualmente chamados de termos.

A terceira etapa consiste na geração da matriz (*tweets* x termos), de acordo com o Modelo Vetorial (*Vector Space Model*) [Baeza-Yates 1999], amplamente utilizado na área de Recuperação de Informações. Nesta pesquisa são geradas 3 matrizes normalizadas, de acordo com o esquema de ponderação empregado para os termos [Tan 2009]: (a) *booleano* (presença ou não do termo no *tweet*), (b) TF (frequência do termo no *tweet*) e (c) TF-IDF (frequência do termo – inverso da frequência de documentos) [Xia 2011] e [Kasyoka 2014], dada para um *tweet d* e um termo *t* por

$$\text{TF-IDF}(d,t) = \text{TF}(d,t).\log(||\text{ND}||/\text{DF}(t)) \tag{1}$$

onde ND é o número total de *tweet* e DF(*t*) é o número de *tweets* na base em que o termo *t* aparece.

Para se obter os vetores correspondentes aos *tweets* basta considerar todos os termos identificados como uma lista ordenada global, de forma que cada termo corresponde a uma dimensão em um espaço NT-dimensional, em que NT é o número total de termos considerados. Normalmente cada *tweet* contém alguns poucos termos e, portanto, o conjunto de mensagens é bastante esparso nesse espaço. Nesta pesquisa a similaridade entre dois *tweet* no espaço NT-dimensional é calculada de duas formas: (a) pela usual distância euclidiana (ou seu inverso que indica similaridade) e (b) pela medida de similaridade do cosseno, que possui valores entre 0 e 1. Estas medidas são dadas correspondentemente pelas fórmulas:

$$euclidiana(d_1,d_2) = \sqrt{\sum_{i=1}^{n}\left(d_1^i - d_2^i\right)^2} \tag{2}$$

$$cosseno(d_1, d_2) =< d_1, d_2 > /||d_1||.||d_2|| \quad (3)$$

onde d_1 e d_2 são representações vetoriais de dois *tweets* e $<,>$ representa o produto interno de dois vetores. Estas duas medidas estão diretamente relacionadas se ambos os vetores são normalizados e contêm apenas valores positivos ou nulos.

Após a etapa de pré-processamento textual aplicam-se os algoritmos de agrupamento. Nesta pesquisa empregaram-se os algoritmos clássicos k-means e k-medoids [Tan 2009], e o recente algoritmo de agrupamento NMF, este empregando duas formas de computação como descrito adiante. Um dos parâmetros utilizados é o número de agrupamentos *k*. Vários valores de *k* foram testados, neste artigo serão apresentados apenas os resultados para k = 3, os demais resultados podem ser encontrados em [Klinczak 2016].

3.2. k-means

A opção mais conhecida para realizar agrupamentos é utilizar o já bem conhecido algoritmo k-means [Han 2001]. Resumidamente, este algoritmo trabalha da seguinte forma: (a) uma série de k pontos iniciais do espaço de características são gerados randomicamente; (b) esses pontos são considerados como centroides dos agrupamentos (ou suas médias); (c) cada vetor correspondente a um *tweet* é usado como entrada, sendo associado ao agrupamento com o centroide mais próximo; (d) o valor do centroide do agrupamento associado ao *tweet* é atualizado de forma a considerar este novo elemento; (e) os passos (c) e (d) são repetidos até que não ocorram mais mudanças de rotulação nos *tweets* e os centroides permaneçam inalterados.

3.3. k-medoids

O algoritmo k-medoids é similar ao algoritmo k-means sendo que neste caso se considera como centroide do agrupamento não a média dos elementos a ele associados, mas sim um ponto central ou "medóide", definido como o vetor pertencente ao conjunto de dados existente mais central (ou mais próximo ao centro). Ou seja, no algoritmo k-means o centroide de um agrupamento é dado pelo vetor médio de seus pontos – não necessariamente um elemento pertencente aos dados, enquanto que no algoritmo k-medoids esses valores são associados a pontos de dados existentes [Han 2001].

3.4. NMF

O método NMF (*Non-Negative Matrix Factorization*) [Lee and Seung 2000] é um método de redução dimensional e um método de agrupamento eficiente [Shinnou 2007] onde se faz a decomposição de uma matriz D em duas matrizes W e H, sendo que as três matrizes são não-negativas, isto é, possuem todos os seus valores positivos ou nulos. Esta não-negatividade torna a interpretação destas matrizes mais simples em diversas aplicações – como no caso do tratamento de textos – e também facilita sua computabilidade. O cálculo de W e H no NMF está associado ao problema de otimização com restrições, como descrito a seguir, e as restrições são impostas pela condição de não-negatividade das matrizes envolvidas.

Na técnica NMF a matriz original *D* (m x n) é decomposta como $D \sim WH$, onde *W* e *H* têm dimensões (m x k) e (k x n), respectivamente, e k é um parâmetro definido pelo usuário que depende da aplicação. No problema tratado por este trabalho k é o número de

agrupamentos e está associado ao número de tópicos a serem encontrados nos *tweets*. A decomposição $D \sim WH$ para dado k na forma geral não possui uma solução analítica, de forma que esta decomposição é obtida de modo aproximado por meios numéricos. Dada uma matriz não negativa *D* (m x n) e um inteiro positivo $p < \min\{m, n\}$ encontram-se duas matrizes não negativas *H* e *W* por meio da minimização da função:

$$f(W, H) = (1/2).||D - W.H||^2 \tag{4}$$

onde *f(W,H)* indica uma norma matricial de [Weisstein 2016], e todos os elementos de *W* e *H* são positivos ou zero, ou seja, $W(i,j) \geq 0$ e $H(i, j) \geq 0$.

Diversos procedimentos podem ser usados para resolver esse problema de otimização, tais como os algoritmos de iteração multiplicativa (*multiplicative update* – MU), algoritmos do gradiente descendente e mínimos quadrados alternantes (*alternate least square* – ALS); esses algoritmos encontram-se resumidos em [e Murray Browne e Amy N. Langville e V. Paul Pauca e Robert J. Plemmons 2006]. O NMF que utiliza a opção de otimização ALS é equivalente a uma forma do algoritmo k-means onde a matriz W contém os centroides e a matriz H contém os indicadores de qual conjunto cada termo pertence [Ding 2005].

Basicamente os dados textuais são associados a valores de ponderação não-negativos, até mesmo porque na maior parte das aplicações, componentes negativos contradizem a realidade física [Pak 2010] e [Naaman 2010]. A fatoração MNF tem sido amplamente utilizada como uma técnica muito útil para utilização em dados de alta dimensão [Pauca 2004] e sendo assim torna-se uma escolha natural para o tratamento de informações textuais [e Murray Browne e Amy N. Langville e V. Paul Pauca e Robert J. Plemmons 2006].

Para comparar os resultados obtidos nos experimentos realizados nesta pesquisa foram empregadas duas métricas: (a) a separação média entre agrupamentos (*between cluster sum of squares* – BSS), e (b) a coesão dos agrupamentos (*within cluster sum of squares* – WSS), dadas respectivamente pelas fórmulas [Jain 1988]:

$$BSS = \sum_{i=1} ||C_i||.(m - m_i)^2 \tag{5}$$

$$WSS = \sum_{i=1} \sum_{x \in C_i} (x – m_i)^2 \tag{6}$$

onde *C* e m indicam respectivamente o tamanho e a média do agrupamento *C* e *m* é a média global do conjunto de dados. Como resultado ideal em um bom agrupamento espera-se um alto BSS e um baixo WSS, pois isso indica que os dados do cada agrupamento estão próximos entre si, e que os agrupamentos distintos estão longe um do outro. Os resultados obtidos são apresentados na próxima seção.

4. Experimentos

De forma a aplicar a metodologia proposta obtiveram-se 3 bases de dados a partir do Twitter no dia 19 de fevereiro de 2016, com contextos iniciais dados pelas *hashtags*

"EVOL", "elNino" e "Zika", obtendo-se 5.000 *tweets* para cada conjunto de dados: (a) EVOL (ou *Evolution Emerging*) é um festival musical que ocorre anualmente em Newcastle, Inglaterra, desde 2002. Em 2016 ele ocorreu em 28 de maio e contou com 40 artistas e dezenas de milhares de expectadores para três dias de festa; (b) o contexto Zika está relacionado à doença já considerada como ameaça em escala mundial, causada pelo vírus ZIKV e transmitida pelo mosquito *Aedes aegypti*, mesmo transmissor da dengue e da febre *Chikungunya*. Esse vírus teve sua primeira aparição em 1947, porém houve contaminação em humanos apenas em 1954 na Nigéria. Atualmente já houve contaminações na Oceania em 2007, na França em 2013 e recentemente no Brasil, em 2015; e (c) o fenômeno meteorológico *El Niño* que ocorre irregularmente em intervalos de 2 a 7 anos, onde os ventos sopram com menos força em todo o centro do Oceano Pacífico, resultando numa acumulação de água mais quente que o normal na costa oeste da América do Sul e, consequentemente, na diminuição da produtividade primária e das populações de peixe. Esse fenômeno tem duração relativamente curta, de 15 a 18 meses, e tem um profundo efeito no clima do hemisfério.

Após o pré-processamento textual as matrizes (*tweets* versus termos) obtidas têm as seguintes dimensões: (a) "EVOL", (5.000 x 2.988); (b) "ElNino", (5.000 x 4.180) e (c) "Zika", (5.000 x 4.973). Todos os experimentos foram realizados utilizando-se as opções de ponderação booleana, TF e TF-IDF, todas com vetores normalizados, e foram empregadas a distância euclidiana e a medida de similaridade do cosseno. Para gerar os agrupamentos foram empregados os algoritmos k-means, k-medoids e NMF com as opções de otimização UM e ALS; como dito anteriormente vários valores k para o número de agrupamentos foram utilizados, porém neste trabalho são apresentados apenas os resultados para k=3.

Apresentam-se na Tabela 1 a seguir os resultados obtidos para as métricas de comparação BSS e WSS em cada caso.

- No caso do contexto inicial "EVOL" verifica-se que o melhor resultado é apresentado pelo algoritmo NMF – MU com a medida de ponderação TF-IDF, tendo um alto valor de BSS e um baixo WSS; as opções distância euclidiana e similaridade do cosseno apresentam resultados similares.
- Para a *hashtag* "Zika" os resultados obtidos são similares, indicando como melhor opção o uso de NMF – MU com o uso da medida de ponderação TF-IDF; novamente o uso da distância euclidiana e da medida do cosseno foi indiferente, apresentando resultados muito próximos.
- Finalmente para a *hashtag* "El Niño" novamente os melhores resultados são obtidos com NMF – MU e TF-IDF, indiferentemente do uso da distância euclidiana ou da similaridade do cosseno.

5. Conclusões e Perspectivas

O objetivo deste trabalho é o de identificar os temas mais discutidos em um recorte das mensagens do Twitter dentro de um contexto inicial relacionando os termos associados a esses temas. Para tanto foram utilizadas diversas opções de pré-processamento textual oriundas da área de Recuperação de Informações, e diversos algoritmos de agrupamento.

No que concerne ao pré-processamento obtiveram-se os termos associados aos *tweets* de acordo com o modelo vetorial empregado no tratamento de textos, e utilizaram-

Table 1. Resultados dos experimentos de agrupamento para as diversas opções de agrupamento, ponderação e medidas de distância, para 3 agrupamentos.

		Evol				Zika				El Niño			
		Euclidiana		Cosseno		Euclidiana		Cosseno		Euclidiana		Cosseno	
		WSS	BSS	WSS	BSS	WSS	BSS	WSS	BSS	WSS	BSS	WSS	BSS
Booleano	k-means	17.283 1.593 1.878	8.474	1.593 18.045 1.038	8.552	4.470 30.538 8.355	1.653	5.433 4.464 33.475	1.645	34.654 4.439 6.295	6.366	35.769 4.488 6.139	5.358
	k-medoids	23.32 18.13 44.66	16.985	23.32 18.13 44.66	16.985	23.345 37.376 7.681	6.394	23.345 37.376 7.681	6.394	25.41 43.06 13.11	10.456	25.41 43.06 13.11	10.456
	NMF Multi-plicative Update	**0,00 235 653**	**24.132**	**753 346 0,00**	**24.132**	**2.783 7.346 4.653**	**2.776**	**4.592 2.998 6.998**	**2.993**	**673 235 209**	**4.976**	**673 235 219**	**5.001**
	NMF Least Square	6.468 4.188 10.422	3.594	6.468 4.188 10.422	3.527	24.467 9.607 8.377	1.592	2.896 5.225 39.001	3.091	26.429 9.317 11.003	2.484	26.429 9.317 11.003	2.484
TF	k-means	18.663 1.910 1.832	8.518	17.690 1.910 2.817	8.507	36.949 4.590 3.350	1.892	36.732 2.511 5.677	1.861	31.854 6.744 9.446	5.354	4.589 3.685 38.732	6.393
	k-medoids	23.769 4.242 16.773	17.715	23.76 49.15 18.16	17.715	24.083 38.170 7.681	6.394	24.083 38.170 7.681	6.394	26.22 43.06 13.11	10.456	26.22 43.06 13.11	10.456
	NMF Multi-plicative Update	**24 2.280 1.095**	**9.279**	**16.452 194 2.679**	**8.346**	**2.347 5.452 5.778**	**2.873**	**5.452 4.347 3.013**	**3.357**	**765 454 0.00**	**5.576**	**582 0.00 922**	**4.236**
	NMF Least Square	4.254 3.280 14.194	4.239	4.254 14.194 3.280	4.239	21.800 13.706 8.622	1.563	6.648 37.551 2.921	3.166	10.529 26.250 11.522	2.486	26.250 10.529 11.522	2.486
TF-IDF	k-means	1.898 1.966 30.477	877	1.966 1.898 30.477	877	29.464 2.700 217	539	29.056 447 2.912	506	642 1.524 24.977	531	1.489 24.452 1.242	492
	k-medoids	19.497 4.000 4.151	11.366	19.497 4.000 4.151	11.366	13.179 4.383 6.408	5.078	13.179 4.383 6.408	5.078	16.121 5.608 7.279	6.179	16.121 5.608 7.279	6.179
	NMF Multi-plicative Update	**456 123 119**	**34.806**	**1.165 0,00 2.595**	**33.2995**	**235 680 233**	**3.326**	**2.673 1.783 997**	**3.278**	**446 27 28**	**4.423**	**446 27 28**	**4.423**
	NMF Least Square	1.937 960 31.081	2.597	1.937 31.081 960	2.597	3.629 7.633 21.103	408	2.128 246 46.104	2.194	1.489 23.937 1.576	662	1.489 23.937 1.576	662

se as medidas de ponderação booleana, TF e TF-IDF, além de duas medidas de similaridade, baseadas na distância euclidiana e na medida de similaridade do cosseno. Quanto ao problema do agrupamento foram analisados os algoritmos k-means, k-medoids e NMF com duas opções de otimização (ALS e UM). Estas variações de processamento foram testadas sobre 3 bases de mensagens obtidas do microblog Twitter a partir de contextos iniciais dados pelas *hashtags* "EVOL", "Zika" e "El Niño". Os resultados obtidos foram comparados quantitativamente utilizando as medidas de coesão (WSS) e de separação (BSS) dos agrupamentos obtidos, sendo que o ideal é obter valores baixos de WSS e altos de BSS, o que ocorre quando os agrupamentos são formados por elementos próximos e quando os agrupamentos estão distantes entre si.

De uma forma geral os resultados obtidos para as 3 bases de dados são similares. O algoritmo NMF empregando a opção de otimização MU – *multiplicative update* apresenta os melhores resultados, tendo também um tempo de execução menor. Uma das grandes vantagens do uso do NMF é que este algoritmo permite que um termo pertença a mais de um agrupamento, permitindo que termos apareçam em assuntos paralelos interligados ao tema principal. Outra diferença significativa é que os conjuntos gerados são relativa-

mente menores que dos gerados pelos outros algoritmos, o que facilita a identificação dos temas. O algoritmo k-means obteve o segundo melhor resultado, porém gera agrupamentos desbalanceados em relação ao número de elementos. No que se refere às medidas de ponderação o esquema TF-IDF apresentou os melhores resultados, seguido pelo uso de TF e pela ponderação booleana. Comprovou-se ainda que a similaridade do cosseno apresenta valores muito próximos aos obtidos com o uso da distância euclidiana, o que pode ser explicado pelo uso de valores normalizados e positivos, como nos experimentos realizados.

Como trabalhos futuros pretende-se dar continuidade a esta pesquisa nas seguintes direções: (a) aplicar a proposta de tratamento a outras redes sociais; (b) aplicar métodos para extração de opiniões dentro dos conjuntos obtidos; (c) estudar a propagação dos temas na rede ao longo da dimensão temporal; (d) aplicar o algoritmo NTF (*non-negative tensor factorization*) quando considerando a dimensão temporal da propagação das mensagens; (e) incluir no estudo da propagação também informações geográficas – como latitude e longitude dos emissores dos *tweets*, permitindo a geração de grafos associados e o estudo da difusão dos temas que aparecem nesse microblog; (f) e finalmente, efetuar testes sobre as interpretações dos resultados por humanos a partir de ferramentas de visualização tais como nuvens de palavras associadas a cada agrupamento.

References

Asur, S. e Huberman, B. A. (2010). Predicting the future with social media. pages 492–499. Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Volume 1.

Baeza-Yates, R. e Ribeiro-Neto, B. (1999). Modern information retrieval.

Ding, C., H. X. e. S. H. D. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. SIAM.

e Murray Browne e Amy N. Langville e V. Paul Pauca e Robert J. Plemmons, M. W. B. (2006). Algorithms and applications for approximate nonnegative matrix factorization. In *Computational Statistics and Data Analysis*, pages 155–173.

Ellison, Nicole B., S. C. e. L. C. (2007). The benefits of facebook "friends:" social capital and college students' use of online social network site. page 1143–1168. Journal of Computer Mediated Communication. Volume 12, Issue 4.

Gupta, A. e Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. Proceedings of the 1st Workshop on Privacy and Security in Online Social Media.

Han, J. e Kamber, M. (2001). Data mining concepts and techniques. Morgan Kaufmann.

Hu, M., L. S. W. F. W. Y. e. S. J. M. (2012). Breaking news on twitter. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Jain, A. K. e Dubes, R. C. (1988). Algorithms for clustering data. Prentice Hall.

Java, Akshay e Song, X. (2007). Why we twitter: Understanding microblogging usage and communities. pages 56–65. Workshop on Web mining and social network analysis.

Kasyoka, P. e Mwangi, W. (2014). A framework for aggregating and retrieving relevant information using tf-idf and term proximity in support of maize production. International Journal of Scientific Technology Research, 3.

Klinczak, Marjori N. M. e Kaestner, C. A. A. (2015). Identification of topics on twitter: Comparison of clustering algorithms and case study. LA-CCI.

Klinczak, M. (2016). Dissertação de mestrado: Identificação e propagação de temas em redes sociais. UTFPR.

Kwak, Haewoon e Lee, C. e. P. H. e. M. S. What is twitter, a social network or a news media? pages 591–600.

Lee, D. D. and Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press.

Lerman, K. e Ghosh, R. (2010). Information contagion: an empirical study of the spread of news on digg and twitter networks. International Conference on Weblogs and Social Media.

Liangjie, Hong e Ovidiu, D. (2011). Predicting popular messages in twitter. Proceedings of the 20th international conference companion on World wide web.

Liu, H. e Gundecha, P. (2012). Mining social media: A brief introduction. Informs.

Naaman, B., L. C.-H. e. B. J. (2010). Is it really about me? message content in social awareness streams. CSCW10.

Nel, F. e Lesot, M. (2011). Information propagation on the web: Data extraction, modeling and simulation. International AAAI Conference on Weblogs and Social Media.

Pak, A. e Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation.

Pauca, V. e Paul Shahnaz, F. (2004). Text mining using non-negative matrix factorizations. SIAM.

Paul, M. J. e Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

Porter, M. F. (1980). An algorithm for suffix stripping. Program.

Shinnou, H. e Sasaki, M. (2007). Refinement of document clustering by using nmf.

Tan, P.-N., S. M. e. K. V. (2009). Introduction to data mining – mineração de dados. Editora Ciência Moderna Ltda. Rio de Janeiro.

Weisstein, E. W. (2016). Frobenius theorem. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/FrobeniusTheorem.html.

Xia, Tian e Chai, Y. (2011). An improvement to tf-idf: Term distribution based term weight algorithm. Journal of Software. Vol. 6 Issue 3.

Xiaohui, Yan, J. G. e. S. L. (2013). Learning topics in short texts by non-negative matrix factorization on term correlation matrix. Proceedings of the 2013 SIAM International Conference on Data Mining.