MAM: Método para Agrupamentos Múltiplos em Redes Sociais Online Baseado em Emoções, Personalidades e Textos

Gustavo Paiva Guedes[1,2], Eduardo Bezerra[1,2], Eduardo Ogasawara[2], Geraldo Xexéo[1]

[1]Programa de Engenharia de Sistemas e Computação – COPPE / UFRJ

[2]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ

{gguedes, ebezerra, eogasawara}@cefet-rj.br, xexeo@cos.ufrj.br

Abstract. *An important problem in social network analysis is the partitioning of its users to discover groups that have common interests or characteristics. Given a collection of objects, typically there is not a single way of clustering. Besides, when objects are users of a social network, each object may be described by several datasets. These datasets offers opportunities to explore users behaviors according to different perspectives. This work describes a multi-view clustering method to cluster objects that contains such properties. Our method produces alternative non-redundant clusterings. Due to their difference, they may reveal novel ways of interpreting these users. We have conducted experiments using a Brazilian online social network named MQD. In MQD users are represented by three datasets. Each one corresponds to a particular perspective: emotion, personality and posts. Our experimental results indicate that our method is able to produce difference clusterings that encompasses the three perspectives of users.*

Resumo. *Um problema importante em análise de redes sociais é o particionamento de seus usuários com o objetivo de descobrir grupos que possuem interesses ou características comuns. Dada uma coleção de objetos, tipicamente não existe apenas uma única maneira de formar as partições. Além disto, quando objetos são usuários de uma rede social, cada objeto pode ser representado por diferentes conjuntos de dados. Esses conjuntos de dados oferecem oportunidades para explorar os comportamentos dos usuários a partir de diferentes perspectivas. Esse trabalho descreve um método agrupamento de múltiplas visões para agrupar objetos que contenham tais propriedades. Os agrupamentos produzidos por nosso método produzem agrupamentos alternativos não-redundantes. Devido a essas diferenças, eles podem revelar novas maneiras de interpretar os dados. Os experimentos conduzidos nesses trabalho usaram uma rede social online brasileira denominada MQD. No MQD os usuários são representados por três conjuntos de dados. Cada um deles corresponde a uma particular perspectiva: emoção, personalidade e postagem. Os resultados experimentais indicam que nosso método é capaz de produzir agrupamentos diferentes que consideram as três perspectivas dos usuários.*

GUEDES, G. P.; BEZERRA, E.; OGASAWARA, E.; XEXEO, G.
MAM: Método para Agrupamentos Múltiplos em Redes Sociais Online Baseado em Emoções, Personalidades e Textos
iSys - Revista Brasileira de Sistemas de Informação, Rio de Janeiro, vol. 7, No. 3, p. 38-55, 2014.

1 Introdução

As redes sociais online se encontram bastante presentes em nossa sociedade. Por meio delas, um usuário pode, por exemplo, compartilhar suas emoções e estados psicológicos com outros usuários. Em redes sociais típicas, cada usuário pode ser representado, por exemplo, pelo conjunto de mensagens que publica, pelos dados de um teste de personalidade que realiza, ou até mesmo pelos padrões de interação com os demais usuários, o que resulta em diversas estruturas de associação, como, por exemplo, a associação de amizade. Nas redes sociais existem alguns aspectos interessantes para estudo, dentre os quais aparecem a detecção de comunidades, predição de links e detecção de padrões. Um problema relevante é agrupar os usuários da rede social com o objetivo de evidenciar padrões associados aos seus comportamentos, características e interesses (Wasserman e Faust 1994) .

Há inúmeras pesquisas recentes em agrupamento de dados que têm mostrado que, dada uma coleção de objetos, há várias maneiras alternativas de agrupá-las, de modo que cada um dos agrupamentos possa revelar uma perspectiva diferente e interessante desses objetos (Bae e Bailey 2006, Davidson e Qi 2008, Xuan Hong Dang 2014). A ideia geral compreende a utilização de algoritmos que possam prover soluções de agrupamentos múltiplos (*multiple clusterings*).

No contexto das redes sociais, embora haja trabalhos que realizam agrupamento com múltiplas visões (Greene e Cunningham 2013), não foram observadas abordagens semi-supervisionadas que combinassem múltiplas visões (*multi-view*) com agrupamentos múltiplos. Isso deixa espaço para investigações dessa lacuna. Tais investigações são relevantes em pesquisas com redes sociais, uma vez que cada usuário pode fazer parte de diferentes grupos a partir de suas diferentes interações na rede. Essas diferentes partições podem ser exploradas por diferentes áreas como, por exemplo, a área de marketing (Dalgic 2006). A Figura 1 ilustra uma rede social hipotética. Agrupar os usuários <[1,2,3,4],[5,6,7]> poderia ser interessante para uma empresa que vendesse roupas para adolescentes. Por outro lado, agrupar os usuários <[1,2,6,7],[3,4,5]> poderia ser mais interessante para uma empresa que vendesse artigos esportivos de futebol. No primeiro caso, a idade poderia ser mais importante e no segundo o gênero.

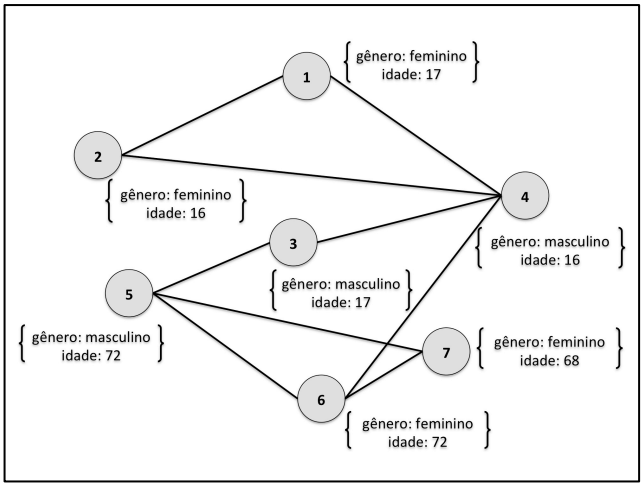

Figura 1. Exemplo de uma pequena rede social.

Este trabalho apresenta o MAM, um método para gerar agrupamentos múltiplos alternativos não-redundantes em uma rede social online a partir de múltiplas visões (emoções, personalidade e textos). O objetivo é gerar agrupamentos alternativos e não-redundantes a partir de uma coleção de usuários representados por dados relacionados a suas emoções, personalidades e postagens. Utilizamos uma abordagem de agrupamento semi-supervisionado, que, comparada à abordagem não-supervisionada oferece o benefício de permitir que informação externa seja incorporada no processo de agrupamento, na forma de relacionamentos (restrições) entre objetos a serem agrupados. Assim, existe a possibilidade de que as restrições reflitam as necessidades dos usuários. Esta abordagem estende a abordagem proposta em Guedes et al. (2013) ao desenvolver um novo método de agrupamento múltiplo, além de apresentar uma avaliação experimental mais completa. Nessa avaliação experimental, utilizamos dados provenientes da rede social online denominada *Meu Querido Diário* (MQD)[1]. O MQD é uma rede social brasileira na qual os usuários descrevem suas experiências diárias, de forma similar ao uso de um diário pessoal. Os resultados obtidos a partir de experimentos computacionais sobre o MQD indicam que o MAM foi capaz de gerar agrupamentos alternativos não-redundantes utilizando diferentes perspectivas dos usuários.

Além dessa introdução, esse trabalho está organizado em mais cinco seções. A seção 2 apresenta uma introdução às técnicas de agrupamento em redes sociais. A seção 3 descreve a utilização de dados sobre emoção, personalidade e postagens em redes sociais. As seções 4 e 5 descrevem, respectivamente, o método proposto e os resultados obtidos. Por fim, a seção 0 descreve a conclusão de nosso estudo.

2 Técnicas de Agrupamento em Redes Sociais

Agrupamento é uma tarefa popular da área de mineração de dados, muitas vezes utilizada como um passo inicial para a análise exploratória de conjuntos de dados complexos. Muitos algoritmos de agrupamento podem ser considerados como procedimentos de otimização discreta orientada por uma função objetivo. Tipicamente, o espaço de busca é bastante grande, posto que cada estado desse espaço corresponde a uma possível partição do conjunto de objetos. O procedimento de otimização tenta encontrar uma partição na qual os objetos de cada grupo sejam semelhantes e objetos diferentes fiquem em grupos distintos (Han et al. 2011).

O resultado final de um agrupamento é uma partição dos dados que apresenta uma perspectiva dos objetos. A maioria dos algoritmos de agrupamento evidencia apenas uma partição dos dados disponíveis (Jain et al. 1999). Entretanto, dados multifacetados têm se tornado relativamente comuns nos últimos anos, o que possibilita a geração de diversas partições não-redundantes dos mesmos dados. Em função disso, diversos algoritmos de agrupamentos múltiplos foram propostos recentemente.

De acordo com Nguyen (2010), as abordagens existentes que tratam sobre agrupamentos múltiplos podem ser divididas em duas categorias: as orientadas a função objetivo e as orientadas a transformação dos dados. Na primeira abordagem, o agrupamento é guiado por uma função objetivo que segmenta os objetos em um número

[1] Disponível em http://www.meuqueridodiario.com.br

pré-estabelecido de grupos. Na abordagem orientada a transformação dos dados, o processo de agrupamento é guiado por uma transformação nos dados antes de se utilizar um algoritmo de agrupamento. Essa transformação tem a intenção de revelar uma nova perspectiva que estava escondida na representação original. Há também pesquisas que investigam algoritmos que utilizam múltiplas visões (*multi-view*) dos dados. Nesse caso, almeja-se produzir um agrupamento que combine os dados dessas visões (Muller et al. 2010).

No contexto de redes sociais, algumas abordagens baseadas em grafos são aplicáveis. Uma das abordagens mais populares é denominada agrupamento espectral (*spectral clustering*), que particiona recursivamente os vértices do grafo que representa a rede social, usando informações estruturais (Luxburg 2007). Recorrentemente, os algoritmos de agrupamento espectral calculam autovalores e autovetores da matriz laplaciana do grafo. Uma das desvantagens desses algoritmos está no custo computacional, que no pior caso é da ordem de $O(n^3)$.

O algoritmo Girvan-Newman (Girvan e Newman 2002) é um método que consiste em remover progressivamente arestas de um grafo para detectar comunidades. Esse método utiliza o conceito de centralidade para encontrar fronteiras entre as comunidades, removendo as arestas com maior centralidade de intermediação. Com isso, a tendência é os componentes que permanecem conectados formem as comunidades. Esse algoritmo possui uma complexidade de $O(n^2)$, embora não consiga produzir agrupamentos múltiplos provenientes de dados multifacetados.

3 Emoção, Personalidade e Postagens em Redes Sociais

Em redes sociais típicas, os usuários podem, por exemplo, escrever postagens (*posts*), realizar comentários e incluir fotos. As redes sociais fornecem um substrato interessante que pode ser analisado por estudiosos de diversas áreas, como sociologia, psicologia e, mais recentemente, ciência da computação. Pennebaker (2013) afirma que diferentes padrões de palavras funcionais (*function words*) revelam partes importantes da personalidade de indivíduos e como eles pensam. Dimitrius e Mazzarella (2008) observam que os padrões de personalidade podem, por exemplo, auxiliar em tomadas de decisão. Neste contexto, aspectos como personalidade e emoções podem ser explorados nessas redes.

A personalidade pode ser definida como um conjunto dinâmico e organizado de características possuídas por uma pessoa que unicamente influencia suas opiniões, motivação e comportamento em várias situações (Ryckman 2013). Na psicologia, a personalidade de um indivíduo é modelada como traços ou fatores. Existem diversos modelos utilizados para realizar essa representação. O modelo dos cinco grandes fatores da personalidade (*Big Five*) se tornou a abordagem dominante para modelar a personalidade na psicologia (Raad e Perugini 2002). Esse modelo é composto pela representação de cinco fatores da personalidade: *abertura*, *conscienciosidade*, *extroversão*, *agradabilidade* e *neuroticismo*. Ao responder a um questionário relacionado aos fatores da personalidade composto por 44 perguntas baseadas na escala de *Likert*, o usuário escolhe um número de 1 a 5 para cada questão, onde 1 denota forte discordância e 5 denota forte concordância. Utilizando as respostas apresentadas a cada uma das 44 perguntas, cada fator da personalidade é calculado da seguinte forma:

- Extroversão: $Q_1 + \overline{Q_6} + Q_{11} + Q_{16} + \overline{Q_{21}} + Q_{26} + \overline{Q_{31}} + Q_{36}$.
- Agradabilidade: $\overline{Q_2} + Q_7 + \overline{Q_{12}} + Q_{17} + Q_{22} + \overline{Q_{27}} + Q_{32} + \overline{Q_{37}} + Q_{42}$
- Conscienciosidade: $Q_3 + \overline{Q_8} + Q_{13} + \overline{Q_{18}} + \overline{Q_{23}} + Q_{28} + Q_{33} + Q_{38} + \overline{Q_{43}}$
- Neuroticismo: $Q_4 + \overline{Q_9} + Q_{14} + Q_{19} + \overline{Q_{24}} + Q_{29} + \overline{Q_{34}} + Q_{39}$
- Abertura: $Q_5 + Q_{10} + Q_{15} + Q_{20} + Q_{25} + Q_{30} + \overline{Q_{35}} + Q_{40} + \overline{Q_{41}} + Q_{44}$

Na lista acima, Q_i denota o valor da i-ésima questão e $\overline{Q_i}$ denota que i-ésima questão é computada com o valor da resposta invertido. Para ilustrar o cálculo dos fatores, considere que um usuário responde às perguntas do teste de personalidade como demonstrado na Tabela 1. Podemos observar que esse usuário marcou 3 para a questão 1, 4 para a questão 6 e assim por diante. Com isso, podemos calcular o fator da *extroversão* conforme a Equação 1. Logo, o usuário apresenta o valor 3,13 representando seu fator de *extroversão*. Os demais fatores da personalidade são calculados de forma análoga.

Tabela 1. Resposta para algumas questões do teste de personalidade.

Questão	Valor
1	3
6	4
11	2
16	3
21	1
26	5
31	4
36	3

$$Extroversão = \frac{(3 + (6 - 4) + 2 + 3 + (6 - 1) + 5 + (6 - 4) + 3}{8} = 3{,}13 \quad (1)$$

No que tange às emoções, essas têm sido estudadas em diversos campos, como psicologia, sociologia e filosofia. Recentemente, pesquisadores a comunidade de ciência da computação tem mostrado interesse por estudos relacionados a emoções, principalmente na área de linguística computacional. Diversas teorias foram propostas para o estudo de emoções, entretanto, a mais frequentemente adotada entre os pesquisadores em processamento de linguagem natural é a proposta (Ekman e Friesen 1978), que propõe a existência de seis emoções básicas: *felicidade*, *tristeza*, *raiva*, *medo*, *nojo* e *surpresa*.

O uso da linguagem de palavras em postagens pode refletir a personalidade, humor, situação social, classe e uma série de outros aspectos sobre os indivíduos (Pennebaker 2002). Neste contexto, uma pergunta interessante a ser respondida no estudo de redes sociais compreende: é possível observar padrões emocionais e de personalidade nas postagens utilizadas para representar os usuários dessas redes? Existem alguns estudos relacionados à personalidade e emoção em redes sociais. Golbeck et al. (Golbeck et al. 2011) demonstra ser possível prever a personalidade de usuários do Twitter utilizando algoritmos de aprendizado de máquina. Esse trabalho também utilizou os cinco traços de personalidade descritos na Seção 3. Wehrli (Wehrli 2008) estuda como as características na personalidade podem influenciar o comportamento nas redes sociais. Nesse contexto, caso a rede social permita que o

usuário escreva textos, disponibilize um teste de personalidade para seus usuários e permita a associação dos textos com emoções, podemos representar os usuários a partir dessas três perspectivas: emoções, personalidades e postagens, trazendo oportunidades para agrupá-los de diferentes formas.

4 MAM - Método de Agrupamento Múltiplo

O Método de Agrupamento Múltiplo (MAM) apresentado nesse trabalho tem o objetivo de agrupar usuários de uma rede social a partir de informações de personalidade, emoções e postagens associadas a esses usuários. Formalmente, cada usuário é representado com uma tripla $(\vec{e}, \vec{p}, \vec{w})$ de vetores nos espaços vetoriais E, P e W, representando, respectivamente, as emoções, personalidades e postagens. Nos próximos parágrafos, descrevemos de que forma esses vetores são formados.

A dimensão E (de emoção) é composta por n valores, para cada usuário. Dessa forma, um usuário u_i é representado em E por um vetor $\vec{e}_i = (e_{i1}, e_{i2}, \ldots, e_{in})$, onde e_{ij} é o valor de entrada associada a emoção j pelo usuário u, tal que $1 \leq j \leq n$. Da mesma forma, representamos cada usuário u no espaço P (de personalidades), composta por m valores, como um vetor $\vec{p}_i = (p_{i1}, p_{i2}, \ldots, p_{im})$. Cada p_{ij} é o valor de entrada associada a personalidade j pelo usuário u, tal que $1 \leq j \leq m$.

Para representar os usuários em W, foi utilizado o modelo de espaço vetorial (Manning et al. 2008). A partir dos conteúdos das postagens de cada usuário do conjunto considerado, é gerado um dicionário, *i.e.*, um conjunto de termos que ocorre ao menos uma vez em pelo menos uma daquelas postagens. Em seguida, são removidas as palavras funcionais (*function words*), (e.g., preposições, artigos, etc.) desse dicionário. Também foi aplicado o processo de *stemming* para reduzir as palavras a sua raiz morfológica.

Considere que T corresponde ao conjunto de termos resultantes do pré-processamento e que |T| = q. Para cada usuário, foi construído o vetor correspondente no espaço W utilizando uma medida conhecida na área de Recuperação de Informação, denominada TF-IDF (Manning et al. 2008). Assim, o usuário u é representado no espaço vetorial W como um vetor $\vec{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iq})$, no qual a componente $w_{ij}, 1 \leq j \leq q$, é computada utilizando a medida TF-IDF. Dado um usuário u_i e o termo t_j, w_{ij} é calculado utilizando a Equação 2. Nesta equação, $|\mathcal{U}|$ corresponde ao número total de usuários, $tf\left(u_i, t_j\right)$ é o número de vezes que um termo t_j ocorre no conjunto de postagens postadas pelo usuário u_i e $df\left(t_j\right)$ é o número de usuários que utilizaram o termo t_j ao menos uma vez em suas postagens. Dessa forma, w_{ij} é um número que reflete o quão importante é um termo t_j nos conteúdos escritos pelo usuário u_i na rede social.

$$w_{ij} = tf\left(u_i, t_j\right) \times idf\left(t_j\right) = tf\left(u_i, t_j\right) \times log\left(\frac{|\mathcal{U}|}{df\left(t_j\right)}\right) \quad (2)$$

As triplas $(\vec{e}, \vec{p}, \vec{w})$ de todos os usuários representam a entrada para nosso método de agrupamento múltiplo. O objetivo é gerar agrupamentos alternativos e não-redundantes a partir de uma coleção de usuários representados pelas suas emoções, personalidades e postagens. Para isso, utilizamos o algoritmo *k-means* (MacQueen 1967). Dada uma coleção de usuários $\mathcal{U}$ e um número k como entrada, o *k-means* gera

partição de $\mathcal{U}$ de tamanho k, isto é, um agrupamento composto por k grupos não sobrepostos. O *k-means* determina k centroides (um para cada grupo) otimizando localmente uma função objetivo que procura maximizar a similaridade dentre de cada grupo e minimizar a similaridade intergrupos.

Nosso método de agrupamento múltiplo é composto por dois passos. O primeiro passo compreende a geração (por meio do *k-means* clássico) de três agrupamentos-base a partir de cada uma das três perspectivas dos usuários (*E*, *P*, *W*). Esses agrupamentos-base são então utilizados para guiar a geração dos demais agrupamentos. O propósito é gerar um agrupamento dos usuários no espaço *W* diferente do agrupamento-base. Para isso, a função objetivo do *k-means* foi modificada para penalizar soluções similares à solução de agrupamento-base. O objetivo dessa modificação é fazer com que as novas soluções sejam distintas do agrupamento-base.

Para formalizar, temos $\mathcal{U} = \{u_i\}$ o conjunto de usuários e cada grupo dos agrupamentos-base possui um rótulo $l \in L$, onde $L = \{1, 2, \ldots, k\}$. Temos $c(u)$: $\mathcal{U} \longrightarrow L$ é a função que retorna o rótulo do grupo associado a um determinado usuário e μ_l é o centroide do grupo C_l. A função objetivo utilizada para gerar os novos agrupamentos distintos do agrupamento-base está definida na Equação 3 (Bezerra et al. 2007, da Bezerra et al. 2006).

$$Obj = \sum_{u_i \in \mathcal{U}} sim\left(u_i, \mu_{c(u_i)}\right) - TotalCostML - TotalCostCL \tag{3}$$

A função objetivo é composta de três parcelas. Na primeira parcela, $sim\left(u_i, \mu_{c(u_i)}\right)$ é a função que calcula a similaridade entre o usuário u_i e o centroide correspondente $\mu_C(u_i)$. A escolha da medida de similaridade depende da visão selecionada para gerar o agrupamento-base. No caso desse trabalho, foi utilizada a distância por cosseno, visto que essa abordagem é comumente adotada para dados esparsos e com muitas dimensões.

A segunda e terceira parcelas da função objetivo correspondem ao custo de violar restrições provenientes dos agrupamentos-base. Considerando u_i e u_j usuários pertencentes ao conjunto de usuários, a notação de uma restrição *must-link* criada entre dois usuários presentes no mesmo grupo pode ser representada por ML(u_i, u_j). Analogamente, representou-se CL(u_i, u_j) como uma restrição *cannot-link*. Estas restrição são criadas de modo a produzir agrupamentos diferentes dos agrupamentos-base. Conforme Bezerra et al. (2007) e da Silva et al. (2006) as Equações 4 e 5 apresentam o custo total de violação de uma restrição *must-link* e *cannot-link*, respectivamente.

$$TotalCostML = \sum_{(u_i, u_j) \in S_{ML}} v_{ij}^{ML} \times \left(1 - I\left(c(u_i), c\left(u_j\right)\right)\right) \tag{4}$$

$$TotalCostCL = \sum_{(u_i,u_j)\in S_{CL}} v_{ij}^{CL} \times I\left(c(u_i), c(u_j)\right) \quad (5)$$

Nas Equações 4 e 5, $I(x, y)$ é a função indicadora (i.e., I retorna 1 quando $x = y$ e I retorna 0 quando $x \neq y$). Os valores de v_{ij}^{ML} e v_{ij}^{CL} são os custos de se violar uma restrição *must-link* e *cannot-link* respectivamente. Vale notar que as regras *TotalCostCL* e *TotalCostML* na função objetivo servem para penalizar as soluções de agrupamento que são similares à solução do agrupamento-base. O número de restrições utilizadas provenientes do agrupamento-base é um parâmetro do nosso método de agrupamento múltiplo.

Para ilustrar o efeito das restrições sobre o processo de agrupamento, considere o exemplo apresentado na Figura 2. Esse exemplo ilustra a utilização de uma restrição ML, supondo que cada usuário seja representado pelas palavras que usou nas postagens que escreveu. As cores azul e vermelho servem apenas para ilustrar adjetivos positivos e negativos. O agrupamento em (b) foi produzido a partir do uso do *k-means* clássico. Ao introduzir uma restrição em (c), uma nova perspectiva pode ser evidenciada em (d), na qual os elementos positivos ficaram em um grupo, enquanto os negativos ficaram em outro. Desta forma, as restrições permitem que sejam formados agrupamentos alternativos para os usuários.

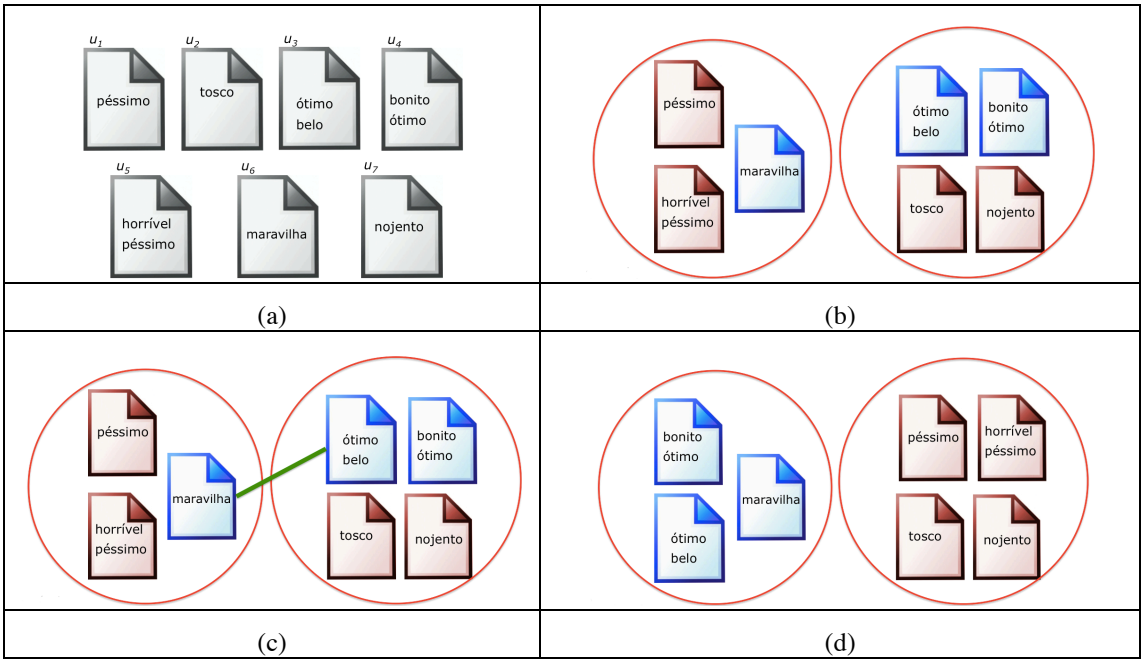

Figura 2. Resultado da aplicação de uma restrição entre os usuários u_3 e u_6: documentos não agrupados em (a); documentos agrupados naturalmente pelo *k-means* em (b); usuários u_3 e u_6 recebendo uma restrição ML em (c); resultado final do agrupamento pelo nosso algoritmo com a restrição ML apresentada em (d).

O pseudocódigo proposto no Método 1 ilustra de forma mais clara o funcionamento da abordagem apresentada nesse trabalho.

Método 1: MAM

1: **Input:**
- s_w= conjunto de dados de postagens
- s_e=conjunto de dados de emoções
- s_p=conjunto de dados de personalidades
- k_w=número de grupos do conjunto de postagens
- k_e =número de grupos do conjunto de emoções
- k_p =número de grupos do conjunto de personalidades
- n_{ML}=número de restrições ML a serem geradas
- n_{CL}=número de restrições CL a serem geradas
- $i$nc=incremento no número de restrições

2: **Output:** C, conjunto de soluções de agrupamentos em s_w.

3: $C \leftarrow \emptyset$

4: $base_w \leftarrow base(s_w, k_w)$

5: $base_e \leftarrow base(s_e, k_e)$

6: $base_p \leftarrow base(s_p, k_p)$

7: $r_{ML} \leftarrow generateMLConstraints(base_p, base_e, n_{ML})$

8: $r_{CL} \leftarrow generateCLConstraints(base_p, base_e, n_{CL})$

9: $\overline{r_{CL}} \leftarrow invert(s_{ML})$

10: $\overline{r_{ML}} \leftarrow invert(s_{CL})$

11: $C \leftarrow generateClusterings(s_w, \overline{r_{CL}}, \overline{r_{ML}}, inc, k_w)$

12: $return\ C$

O método inicia recebendo nove parâmetros. Em seguida, nos passos 4, 5 e 6, os agrupamentos-base para cada um dos conjuntos de dados (emoções, personalidades e postagens) são obtidos a partir do *k-means* clássico utilizando um número *k* de grupos.

No passo 7, a função *generateMLConstraints* gera um número n_{ML} de restrições *must-link* utilizando os agrupamentos $base_e$ e $base_p$. Primeiramente, a função seleciona os pares de usuários pertencentes ao mesmo grupo tanto em $base_e$ como em $base_p$ para gerar restrições ML. Com isso, se dois usuários u_i e u_j pertencem ao mesmo grupo em $base_e$ e pertencem ao mesmo grupo em $base_p$, foi gerada uma restrição ML. Assim, geramos todas as restrições ML possíveis. Em seguida, essa função calcula a distância euclidiana entre esses pares de objeto e atribui as n_{ML} restrições mais similares a r_{ML}. Analogamente, o passo 8 realiza o procedimento inverso para produzir as restrições CL. Nesse caso, são selecionados os pares de usuários que estão em grupos distintos tanto em $base_e$ como em $base_p$ e as n_{CL} restrições mais similares são atribuídas a r_{CL}.

Os passos 9 e 10 utilizam a função *invert* para inverter todas as restrições presentes em r_{ML} e r_{CL}. Assim, as restrições *must-link* presentes em r_{ML} são

transformadas em restrições *cannot-link* e inseridas em $\overline{r_{CL}}$. Analogamente, as restrições *cannot-link* presentes em r_{CL} são transformadas em restrições *must-link* e inseridas em $\overline{r_{ML}}$.

A função *generateClusterings*, invocada no passo 11, é responsável por aplicar, separadamente, as restrições $\overline{r_{CL}}$ e $\overline{r_{ML}}$ no conjunto de dados s_w. O retorno da função *generateClusterings* é um conjunto de agrupamentos C. Os agrupamentos gerados a partir do conjunto de restrições $\overline{r_{CL}}$ são gerados com 0 a $|\overline{r_{CL}}|$ restrições, incrementados com o valor de inc. Assim, se tivermos o valor $inc = 50$ e $|\overline{r_{CL}}| = 100$, teremos 3 agrupamentos, o primeiro gerado com 0 restrições, o segundo com 50 e o último com 100. Da mesma forma, os agrupamentos gerados a partir de $\overline{r_{ML}}$ são gerados com 0 a $|\overline{r_{ML}}|$ restrições, incrementados com o valor de inc. Com a utilização dessas restrições e a função objetivo modificada do *k-means* (Equação 3), o propósito é produzir agrupamentos distintos dos agrupamentos-base.

5 Avaliação Experimental

5.1 Conjunto de dados

No presente trabalho, utilizamos dados provenientes de uma rede social online chamada Meu querido Diário (MQD). Essa rede permite que usuários escrevam postagens e associem *marcações de emoções* a elas. Durante a utilização do MQD, os usuários podem descrever o que fizeram durante o dia, quais seus sentimentos ou alguma informação sobre seus estados emocionais. Além disso, podem escolher uma entre seis emoções para associar a suas postagens. Essas emoções fazem parte das *seis emoções básicas* propostas por Ekman e Friesen (Ekman e Friesen 1978).

Os usuários do MQD também podem responder a um teste de personalidade (Andrade 2008), que é uma versão em português do Brasil do Modelo dos Cinco Fatores da Personalidade proposto por Piedmont (2008). As informações presentes no conjunto de dados do MQD também apresentam idade, sexo, data de nascimento, estado, estado civil, dentre outros. Quando o usuário escreve uma postagem, ele necessita inserir o título, texto e data do evento. Caso o usuário queira, pode associar sua postagem a uma emoção, mas isso não é obrigatório. Assim como em outras redes sociais online, os usuários podem escrever comentários em cada postagem. Cada postagem do MQD pode ter diversos comentários.

Atualmente, existem mais de 47.000 usuários cadastrados e mais de 11.000 responderam ao teste de personalidade. Além disso, o MQD possui mais de 26.000 relações de amizade e aproximadamente 51.000 postagens com emoções associadas. Existem aproximadamente 100.000 comentários escritos. Para esse estudo, foi utilizado um conjunto de dados no qual todos os usuários responderam ao teste de personalidade e escreveram ao menos cinco postagens com emoções associadas. Essa base foi denominada MQD1093 e possui 1.093 usuários e 20.047 postagens com emoções associadas.

É importante ressaltar que existem três perspectivas sobre os mesmos usuários: emoções, personalidades e postagens. Dado que o objetivo principal do website a partir do qual o conjunto de dados MQD1093 ("MQD1093") foi gerado é a escrita de postagens (a realização do teste de personalidade e a associação de emoções a cada

postagem são opcionais), nosso método consiste em aplicar restrições apenas no conjunto de dados de postagens.

5.2 Metodologia

O objetivo do MAM é utilizar as restrições geradas a partir dos agrupamentos-base de emoções e personalidades para gerar um conjunto C de agrupamentos alternativos não-redundantes sobre o conjunto de postagens . Em seguida, cada agrupamento C_i é comparado com o agrupamento base de postagens $base_w$, de forma que se possa avaliar a redundância desses agrupamentos. Essa avaliação é feita com a utilização de algumas medidas de qualidade, conforme mencionado na Seção 1.

Há dois tipos de medidas para avaliação da qualidade de um agrupamento disponíveis na literatura: as internas e as externas. Na avaliação interna, a própria função-objetivo empregada no agrupamento é usada como medida de qualidade. O propósito é que os valores da função objetivo para os agrupamentos C_i fiquem próximos ao valor da função objetivo do agrupamento $base_w$.

As medidas externas de validação comparam o resultado obtido por um algoritmo de agrupamento com um "*gold standard*", ou seja, um conjunto de dados em que os rótulos são conhecidos e cada objeto pertence a apenas um grupo. Utilizamos duas medidas externas para a avaliação da qualidade dos agrupamentos gerados: Pureza e NMI. Essas medidas são empregadas com o intuito de avaliar se os agrupamentos gerados C são distintos de $base_w$. Nesse caso, $base_w$ é considerado o "*gold standard*".

O índice de pureza de um agrupamento é dado pela soma ponderada da pureza de cada grupo. Quanto mais semelhantes são dois agrupamentos, mais a medida se aproxima de 1 ao passo que quanto mais distintos os agrupamentos, mais a medida se aproxima de 0. Essa medida calcula a relação entre a classe dominante e o tamanho do grupo. Considere que $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ é o conjunto de grupos e $\mathbb{C} = \{c_1, c_2, \dots, c_j\}$ representa o conjunto de classes. Cada elemento de Ω é o rótulo do usuário u_l nos novos agrupamentos e cada elemento de $\mathbb{C}$ é o rótulo do usuário u_l no agrupamento-base. A Eq. (6) apresenta a expressão utilizada para o cálculo do índice de pureza de um agrupamento Ω.

$$purity(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} \max_{j} \left| \omega_k \cap c_j \right| \tag{6}$$

No caso extremo em que cada objeto representa um grupo, seu valor é 1, pois a classe dominante sempre será a classe do único objeto. Da mesma forma, essa medida não avalia os demais elementos do grupo; apenas considera os elementos da classe de maior ocorrência, não avaliando se os demais objetos são todos de uma classe ou de classes variadas. Por essa razão, consideramos a inclusão do índice NMI.

O NMI é uma medida proveniente da Teoria da Informação capaz de quantificar a informação comum entre duas distribuições, que no nosso caso, dois agrupamentos. Essa medida, diferente da Pureza, não considera apenas os elementos da classe de maior ocorrência. A Equação 7 representa a medida do NMI, onde, I corresponde à informação mútua, apresentada na Equação 8. A informação mútua calcula a quantidade de informação que a presença ou ausência de um objeto contribui para a classificação

correta em um grupo. $P(\omega_k)$ é a probabilidade de um objeto pertencer ao grupo ω_k, $P(c_j)$ é a probabilidade de um objeto pertencer à classe c_j e $P(\omega_k \cap c_j)$ é a probabilidade de um objeto pertencer a um grupo ω_k e a uma classe c_j ao mesmo tempo.

$$NMI(\Omega, \mathbb{C}) = \frac{I(\Omega, \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2} \quad (7)$$

$$I(\Omega, \mathbb{C}) = \sum_k \sum_j P(\omega_k \cap c_j) \log \frac{P(\omega_k \cap c_j)}{P(\omega_k)P(c_j)} \quad (8)$$

A informação mútua possui um problema análogo ao encontrada na pureza. A utilização do denominador $[H(\Omega) + H(\mathbb{C})]/2$ normaliza a informação mútua, resolvendo esse problema, onde H é a entropia, demonstrada nas Equações 9 e 10. A entropia tende a ser maior conforme haja aumento no número de grupos. Essa medida calcula a incerteza em uma variável aleatória. Dessa forma, quão maior for a incerteza, maior será o valor da entropia.

$$H(\Omega) = -\sum_k P(\omega_k) \log P(\omega_k) \quad (9)$$

$$H(\mathbb{C}) = -\sum_j P(c_j) \log P(c_j) \quad (10)$$

5.3 Resultados

Nessa seção apresentamos dois experimentos realizados com o método MAM. O primeiro considerou o uso das restrições CL e ML, variando-se o número de restrições de 0 a 500 com incrementos de 50. O segundo considerou apenas as restrições ML, variando-se de 0 a 50 com incremento de 1. Os agrupamentos produzidos pelo MAM são comparados com o agrupamento-base aplicado diretamente sobre os dados.

O agrupamento-base para cada um dos conjuntos de dados (emoções, personalidades e postagens) foi obtido a partir do *k-means* clássico. O critério para escolha do valor de *k* para cada um dos conjuntos de dados utilizou o ponto máximo de curvatura (Munaga et al. 2012), considerando-se agrupamentos variando entre 2 a 20. Como resultado, o número de grupos em cada perspectiva foi definido da seguinte forma: emoções (k=6), personalidades (k=4), postagens (k=8).

Como o conjunto de postagens apresentou o valor de k=8, optamos por apresentar os resultados com uma variação no número *k* de grupos das postagens entre 6 e 10, de forma que pudéssemos observar a variação no comportamento dos novos agrupamentos. Os experimentos não consideraram variações nos valores de *k* para emoções e personalidades, pois os mesmos são utilizados apenas para a geração das restrições a serem aplicadas no conjunto de postagens.

Conforme descrito na Seção 5.2, as medidas de NMI e Pureza foram utilizadas para avaliar a qualidade dos agrupamentos, assim como a própria função objetivo. A Figura 1(a) ilustra que a medida NMI decresce conforme o número de restrições CL aumenta, evidenciando que, de forma geral, quanto mais restrições, mais os novos agrupamentos de postagens gerados se diferenciam de $base_w$. Da mesma forma, a

Figura 1(b) demonstra que a Pureza também tende a decrescer em função do aumento do número de restrições CL. Para exemplificar, podemos observar o valor de *k=7*. Com 500 restrições, o valor do NMI se aproxima de 0.5, indicando que a qualidade do agrupamento gerado é alta.

A Figura 1(c) ilustra a função objetivo dos agrupamentos gerados. Embora a função objetivo cresça conforme o número de restrições inseridas aumente, esse crescimento não é significativamente relevante, o que se pode ser analisado a partir da tabela 2. Podemos verificar que a função objetivo teve uma variação de menos de 1% para todos os novos agrupamentos gerados, quando comparados aos agrupamentos-base de palavras. Isso indica que conseguimos um resultado relevante, visto que, por mais que tenha havido uma variação de 1% na função objetivo, o NMI e a Pureza variaram de forma considerável (e.g. 50.30% para *k=7*).

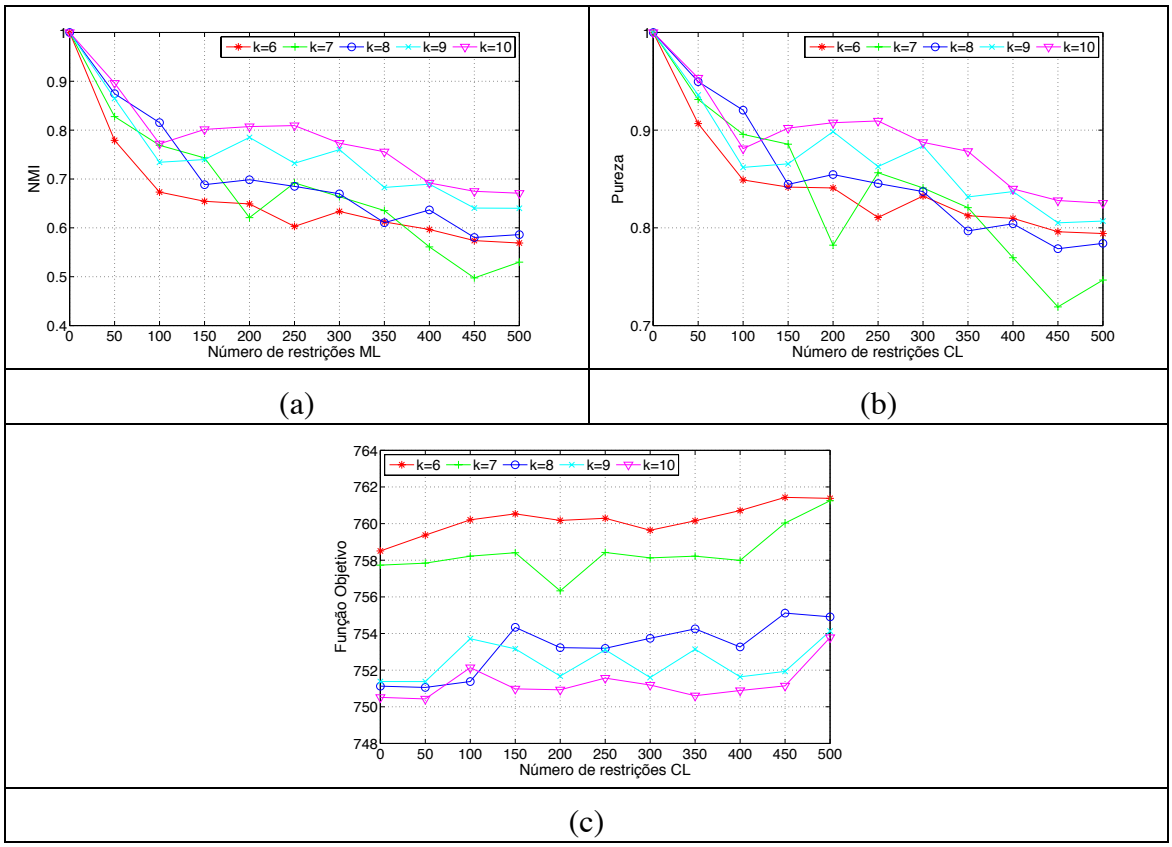

Figura 1. Resultado da aplicação de restrições CL: em NMI (a) em função objetivo (b).

A Tabela 2 apresenta a comparação entre os agrupamentos gerados pelo MAM com restrições CL e o agrupamento-base de postagens. Os números apresentados representam a variação percentual entre o máximo e mínimo para a função objetivo, pureza e NMI nos diferentes valores de *k*. Como exemplo, podemos observar que para *k=6* houve uma variação percentual de 0,38% entre a não-utilização de restrições (0 restrições) e a utilização de 500 restrições.

Tabela 2. Porcentagem de variação na mudança da pureza, NMI e função objetivo para restrições CL.

K	Função Objetivo [%]	Pureza [%]	NMI [%]
6	0,38	20,60	43,11
7	0,46	23,06	50,30
8	0,54	22,15	41,98
9	0,36	19,31	36,00
10	0,44	17,48	32,92

Os resultados apresentados pelas restrições ML apresentaram uma diferença significativa com relação às restrições CL. Ao analisar a Figura 3(a), pode-se observar que a inclusão de um pequeno número de restrições provoca uma alteração nos resultados da função NMI, fazendo com que a mesma alcance um resultado menor que 0,2 com apenas 50 restrições (para todos os valores de k). Da mesma forma, a Figura 3(b) ilustra que a Pureza apresenta uma queda bastante relevante com a inclusão de aproximadamente 50 restrições. A pequena variação na função objetivo pode ser observada na Figura 3(c), que evidencia que existe uma tendência de aumento da função objetivo conforme se aumenta o número de restrições.

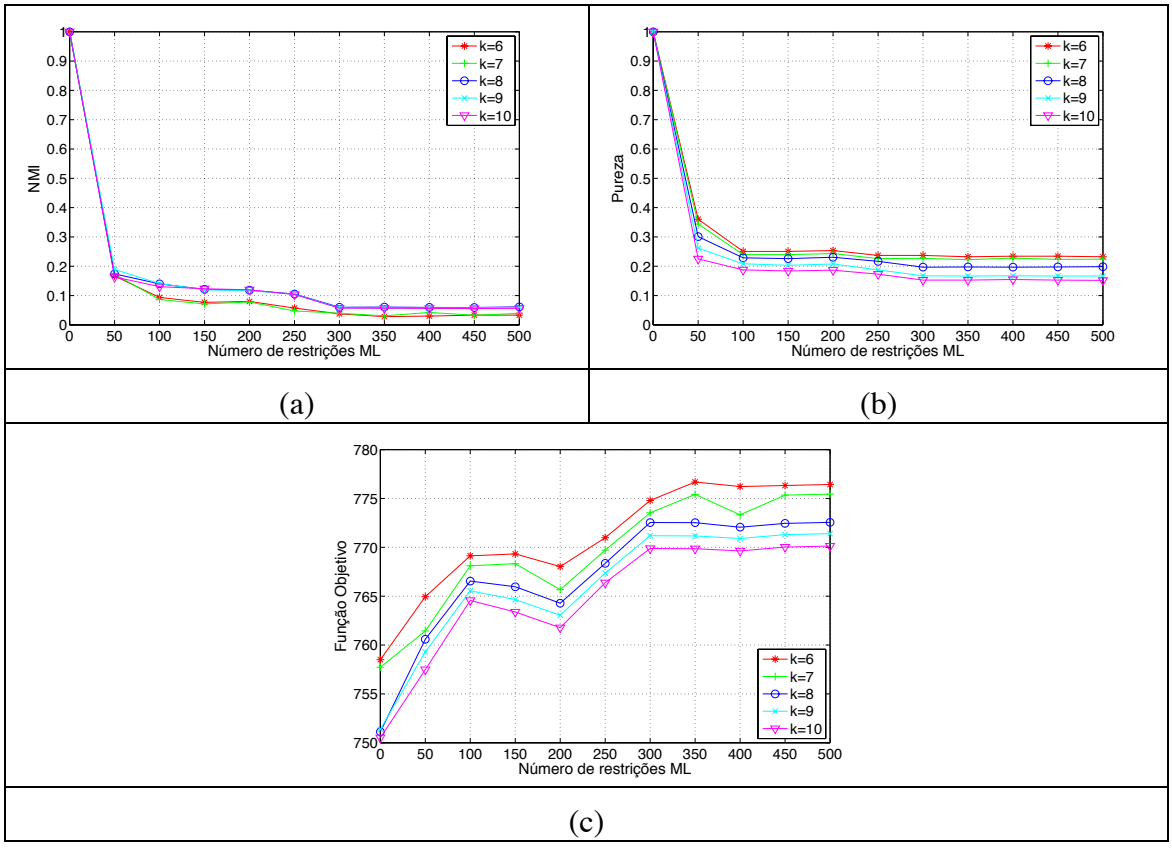

Figura 3. Resultado da aplicação de restrições ML: em NMI (a) em Pureza (b) em função objetivo (c).

A Tabela 3 apresenta a comparação entre os agrupamentos gerados pelo MAM utilizando restrições ML e o agrupamento-base de postagens. Os números apresentados representam a variação percentual entre o máximo e mínimo para a função objetivo, pureza e NMI nos diferentes valores de k.

Tabela 3. Porcentagem de variação na mudança da pureza, NMI e função objetivo para restrições ML.

K	Função Objetivo [%]	Pureza [%]	NMI [%]
6	2.,2	76,80	96,70
7	2,28	77,80	96,92
8	2,77	80,33	94,06
9	2,59	83,35	94,31
10	2,55	84,82	94,44

Ressalta-se que houve uma mudança relativamente pequena na função objetivo (menor que 3%) quando comparada às funções objetivo dos demais agrupamentos (incluindo o agrupamento-base). Por outro lado, observamos uma alta percentagem nas medidas de Pureza e NMI (em torno de 80% e 95% respectivamente), o que indica um bom resultado. A utilização de restrições ML apresentou um comportamento distinto do apresentado pelas restrições CL, pois com um número relativamente pequeno de restrições (50) foi alcançado um valor muito baixo nas medidas externas de qualidade, significando que a geração de agrupamentos com restrições ML necessitam de poucas restrições para se tornarem muito distintos do agrupamento-base. Assim, foi realizado um novo experimento utilizando um número de restrições de 0 a 50 com variações de uma em uma conforme ilustra a Figura 4. Pode-se observar que com um número entre cinco e dez restrições, os valores mínimos da medida de NMI foram alcançados. Os valores apresentados para a Pureza e o NMI destacam que os novos agrupamentos formados variaram significantemente com relação ao agrupamento-base.

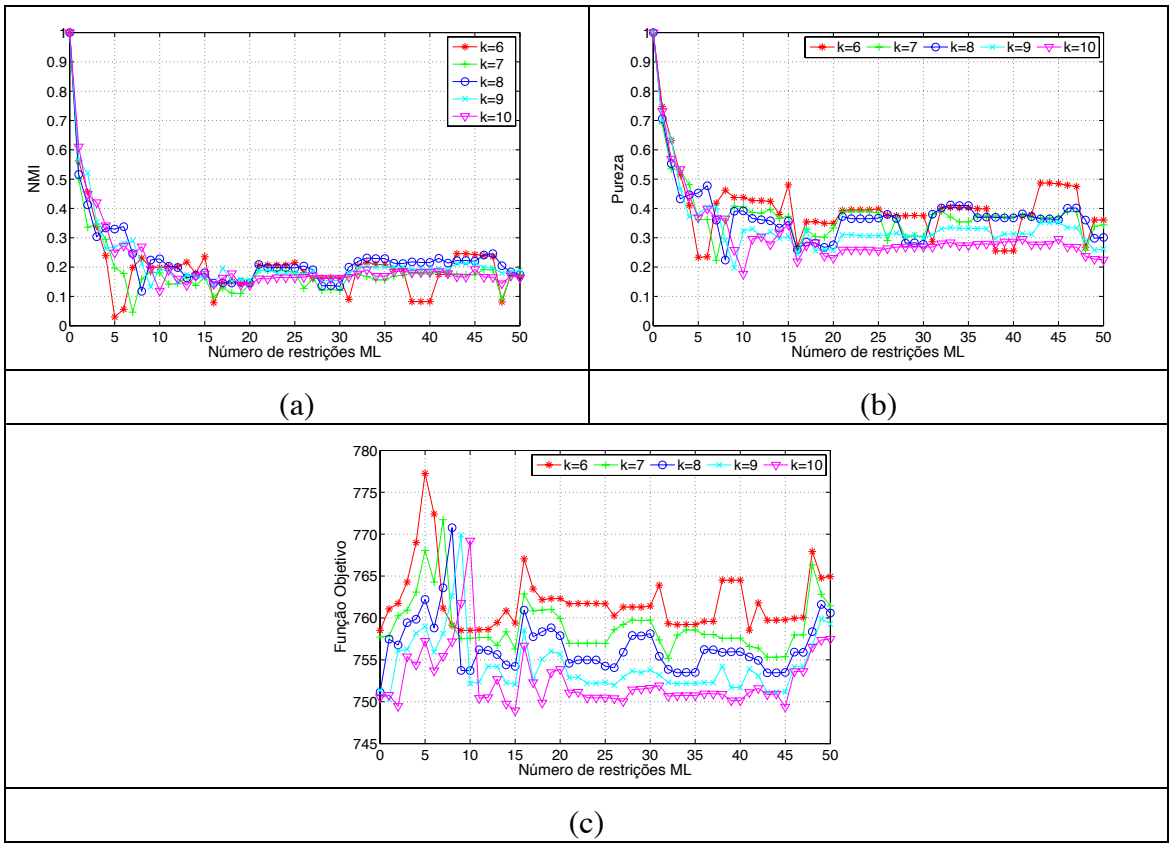

Figure 4. Resultado da aplicação de restrições ML: em NMI (a) em função objetivo (b).

A tabela 4 apresenta o número de restrições ML variando de 1 a 50. A utilização dessas restrições geraram alta variação nas medidas de pureza e NMI, enquanto a função objetivo variou menos de 4% em todos os casos, o que demonstra um resultado bastante satisfatório.

Tabela 4. Porcentagem de variação na mudança da pureza, NMI e função objetivo para restrições ML-1-50.

K	Função Objetivo [%]	Pureza [%]	NMI [%]
6	2.41	76.67	97.05
7	2.14	77.68	95.40
8	2.55	77.59	88.27
9	2.55	81.37	86.72
10	3.03	82.35	88.13

6 Conclusão

Nesse estudo, apresentamos o MAM, uma nova abordagem para gerar agrupamentos múltiplos em redes sociais online baseadas em emoções, personalidade e textos. Nosso algoritmo permite a entrada de restrições, o que faz com que soluções semelhantes aos agrupamentos gerados naturalmente sejam penalizadas. A utilização de uma abordagem semi-supervisionada permite que o usuário final possa interferir na geração dos novos agrupamentos.

Em nossa avaliação, utilizamos restrições provenientes de perspectivas distintas para cada usuário, baseadas em emoções e personalidades. As restrições foram selecionadas quanto maior fosse a similaridade entre os pares de usuários. Em seguida, essas restrições foram aplicadas na perspectiva de postagens. A abordagem foi implementada tomando como base uma versão modificada do algoritmo *k-means* que permite incluir restrições com intenção de gerar soluções alternativas às obtidas naturalmente pelo algoritmo do *k-means*. As restrições indicam que dois objetos devem ficar no mesmo grupo (*must-link*) e que não devem ficar no mesmo grupo (*cannot-link*).

Os experimentos realizados utilizaram os dados extraídos da rede social online MQD. Os experimentos variaram o número de grupos de entrada para o método proposto (de 6 a 10). Para avaliar nosso experimento, foram utilizadas três medidas de qualidade: Pureza, NMI e a função objetivo. Essas medidas demonstraram resultados significantes visto que foi possível gerar agrupamentos alternativos não-redundantes quando comparados ao agrupamento-base, havendo apenas uma pequena variação na função objetivo: menos de 6% em todo o experimento com as restrições ML e CL. As medidas externas de avaliação (NMI e Pureza) apresentaram valores baixos, significando que agrupamentos alternativos e não-redundantes foram gerados. A medida interna de avaliação (função objetivo) apresentou uma variação pequena, indicando que o resultado é bastante satisfatório. Nesse contexto, foi possível gerar agrupamentos alternativos com boa qualidade.

Observamos alguns trabalhos futuros que podem derivar da abordagem proposta nesse trabalho. Dentre eles, a criação de um novo método capaz de lidar com perspectivas genéricas (não apenas com emoções, personalidades e postagens), bem

como a utilização de métricas da estrutura das redes sociais para geração dos agrupamentos, como, por exemplo, a centralidade ou o grau. Um outro trabalho interessante, seria um estudo comparativo entre o MAM e outros trabalhos da literatura. Além disso, planejamos investigar a correlação entre as visões, visto que pode haver dependências entre elas.

Referências

Andrade, J. M. de, (2008). Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil. Disponível em: http://repositorio.unb.br/handle/10482/1751. Acesso em: 1 abr 2014.

Bae, E., Bailey, J., (2006), "COALA: A Novel Approach for the Extraction of an Alternate Clustering of High Quality and High Dissimilarity". In: *Sixth International Conference on Data Mining, 2006. ICDM '06*, p. 53–62

Bezerra, E., Xexéo, G., Mattoso, Marta, (2007), "On the Usage of Structural Information in Constrained Semi-Supervised Clustering of XML Documents", *Successes and New Directions in Data Mining:*, IGI Global

Dalgic, T., (2006), *Handbook of Niche Marketing: Principles and Practice*. Best Business Books, Haworth Reference Press.

Davidson, I., Qi, Z., (2008), "Finding Alternative Clusterings Using Constraints". In: *Eighth IEEE International Conference on Data Mining, 2008. ICDM '08*, p. 773–778

Dimitrius, J.-E., Mazzarella, M., (2008), *Reading people: how to understand people and predict their behavior-- anytime, anyplace*. New York, Ballantine Books.

Ekman, P., Friesen, W., (1978), *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.

Girvan, M., Newman, M. E. J., (2002), "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences*, v. 99, n. 12 (nov.), p. 7821–7826.

Golbeck, J., Robles, C., Edmondson, M., Turner, K., (2011), "Predicting Personality from Twitter". In: *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, p. 149–156

Greene, D., Cunningham, P., (2013), "Producing a Unified Graph Representation from Multiple Social Network Views". In: *Proceedings of the 5th Annual ACM Web Science Conference*, p. 118–121, New York, NY, USA.

Guedes, G., Bezerra, E., Geraldo Xexéo, (2013), "Multi-view Clustering in a Social Network".

Han, J., Kamber, M., Pei, J., (2011), *Data Mining: Concepts and Techniques, Third Edition*. 3 ed. Morgan Kaufmann.

Jain, A. K., Murty, M. N., Flynn, P. J., (1999), "Data clustering: a review", *ACM Comput. Surv.*, v. 31, n. 3, p. 264–323.

Luxburg, U. von, (2007), "A tutorial on spectral clustering", *Statistics and Computing*, v. 17, n. 4 (dez.), p. 395–416.

MacQueen, J. B., (1967), "Some Methods for Classification and Analysis of MultiVariate Observations". In: *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, p. 281–297

Manning, C. D., Raghavan, P., Schütze, H., (2008), *Introduction to Information Retrieval*. Cambridge University Press.

Muller, E., Gunnemann, S., Färber, I., Seidl, T., (2010), "Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data". In: *2010 IEEE 10th International Conference on Data Mining (ICDM)*, p. 1220–1220

Munaga, H., D. R. Mounica Sree, M., V. R. Murthy, J., (2012), "DenTrac: A Density based Trajectory Clustering Tool", *International Journal of Computer Applications*, v. 41, n. 10 (mar.), p. 17–21.

Nguyen, J. E. X. V., (2010), "minCEntropy: A Novel Information Theoretic Approach for the Generation of Alternative Clusterings.", p. 521–530.

Pennebaker, J. W., (2002), "What our words can say about us: Toward a broader language psychology", *Psychological Science Agenda*, v. 15, n. 1, p. 8–9.

Pennebaker, J. W., (2013), *The secret life of pronouns: what our words say about us*. New York, Bloomsbury Press.

Piedmont, R. L., (2008), "The revised NEO Personality Inventory: Clinical and research applications"

Ryckman, R. M., (2013), *Theories of personality*. Australia; Belmont, CA, Wadworth Cengage Learning.

Da Silva, E. B., Mattoso, M., Xexéo, G., (2006), "Semi-Supervised Clustering of XML Documents: Getting the Most from Structural Information.". In: *ICDE Workshops*, p. 88

Wasserman, S., Faust, K., (1994), *Social Network Analysis: Methods and Applications*. 1 edition ed. Cambridge ; New York, Cambridge University Press.

Wehrli, S., (2008), *Personality on Social Network Sites: An Application of the Five Factor Model*, ETH Zurich Sociology Working Paper 7, ETH Zurich, Chair of Sociology. Disponível em: http://econpapers.repec.org/paper/etswpaper/7.htm.

Xuan Hong Dang, J. B., (2014), "Generating multiple alternative clusterings via globally optimal subspaces", *Data Mining and Knowledge Discovery*

Big Five Assessment. , (2002), 1st edition ed. Seattle, WA, Hogrefe & Huber Pub.

.MQD1093. Disponível em: http://sourceforge.net/p/gpca/wiki/MQD1093/. Acesso em: 7 nov 2014.