Recomendação de Grupos de Facebook Baseado em Textos e Imagens de Usuários

Gustavo Ramos Lima
Universidade Federal do Espírito Santo
Av. Fernando Ferrari, 514, Vitória-ES, Brasil
grlneutron@hotmail.com

Patrick Marques Ciarelli
Universidade Federal do Espírito Santo
Av. Fernando Ferrari, 514, Vitória-ES, Brasil
patrick.ciarelli@ufes.br

RESUMO

Este artigo tem como foco a recomendação de grupos do Facebook a usuários, baseado no perfil de postagem de cada usuário no Facebook. Para realizar esta tarefa, os textos e as imagens divulgadas pelos usuários são usados como fonte de informação, e os experimentos indicaram que o uso dos dois tipos de informações juntas fornece resultados melhores ou iguais ao uso das mesmas informações de forma separada. Este trabalho foi resultado de uma iniciação científica, na qual o aluno realizou as tarefas de revisão de literatura, pré-processamento das bases de dados, os experimentos de classificação, implementou parte do código e usou códigos de terceiros, conforme indicado no artigo. O aluno também contribuiu parcialmente na escrita deste artigo.

PALAVRAS CHAVE. Recuperação da Informação, Redes Sociais, Classificação.

DEA, OA

ABSTRACT

This article focuses on the recommendation of Facebook groups to users, based on the post profile of each user on Facebook. In order to accomplish this task, texts and images provided by users are used as a source of information, and the experiments showed that the combination of these two types of information gives results better than or equal to the results obtained when using separately these data. This work was result of a scientific initiation, in which the student performed literature review, preprocessing of data sets, the classification experiments, implemented part of the code and used third-party codes, as indicated in article. The student also contributed partially to write this article.

KEYWORDS. Information Retrieval. Social Networks. Classification.

DEA, OA

1. Introdução

O crescimento das redes sociais online, como Facebook, WhatsApp, Twitter e Instagrama, tem sido imenso na última década, graças ao acesso cada vez mais fácil a internet e ao barateamento de computadores e dispositivos móveis que permitem conexão a essas redes. Atualmente, existem mais de 2 bilhões de usuários no mundo conectados em ao menos uma rede social e a perspectiva é que este número aumente a cada ano [Statista, 2016].

Várias redes sociais e muitos usuários deixam as mensagens, imagens e vídeos para acesso público, podendo qualquer pessoa ler o ser conteúdo e, inclusive, interagir com o usuário que a divulgou. Somado ao fato que existe uma grande quantidade de mensagens divulgadas por minuto nas redes sociais (só para o Facebook é estimado em cerca de 4 milhões de posts por minuto [WERSM, 2015]), isto significa que a quantidade de dados disponíveis não só é gigantesca mais também abrange grande parte da população mundial.

Com esses dados é possível identificar hábitos, interesses, padrões de comportamento e outras informações sobre um usuário que podem ser utilizados para modelar o seu perfil. Com perfis de usuários é possível oferecer uma variedade de aplicações, tais como mecanismos personalizados de busca, recomendação de amigos e conteúdo, identificação de usuários com comportamento malicioso e indicação de produtos e serviços direcionados aos interesses e necessidades dos usuários [Kosinski et al., 2013; Jin et al., 2013; Mislove et al., 2010]. Na área de marketing, é cada vez mais comum as empresas buscarem saber mais sobre os seus consumidores de forma a oferecer produtos de interesse e elevar seus lucros. Abordagens desta natureza já são usadas por grandes empresas, como a Amazon e o Submarino. Para além dos usos comerciais, pesquisas científico-acadêmicas contribuem igualmente através da exploração de múltiplos processos, colocando em questão métodos e propondo análises diversificadas.

Uma abordagem comum para identificar informações relevantes dos usuários é através dos textos que são publicados nas redes sociais, pois são muito informativos e é o meio primário de comunicação nas maioria da redes. No entanto, existem alguns fatos que dificultam a extração de informação dos textos, alguns deles são os erros de ortografia, palavras com vários significados semânticos, sinônimos, os diversos idiomas no mundo e o imenso dicionário de palavras que cada idioma possui. Uma opção para obter mais informações dos usuários, e tornar os perfis mais precisos, é utilizar dados provenientes de imagens e fotografias. Essa opção é viável pois é muito comum usuários não só divulgarem mensagens como também imagens nas redes sociais. Em [Alqhtani et al., 2015] é utilizada uma abordagem que combina as informações de texto e imagem, e foi obtida uma ligeira melhora na qualidade dos resultados quando combinado esses dois tipos informações em comparação ao usá-las separadamente. Esta observação motiva uma melhor análise da combinação desses dados.

Neste contexto, este trabalho analisa o uso de textos e imagens na tarefa de recomendar grupos de Facebook para usuários. Para realizar esta tarefa, foi utilizada uma base de dados capturada de grupos de Facebook, do qual foram selecionados alguns dos usuários com maior frequência de postagem nestes grupos. Baseado no histórico de postagens realizadas pelos usuários e numa modelagem de cada grupo, são recomendados os grupos os quais as pessoas têm maior afinidade. As informações foram extraídas tanto dos textos quanto das imagens, sendo testado três tipos de abordagens: uma usando somente informação textual, outra usando informação das imagens e uma última combinando ambas informações. Os resultados obtidos indicaram que o uso dos dois tipos de dados combinados pode fornecer resultados melhores do que usando eles de forma separada.

Este trabalho possui a seguinte estrutura. Na Seção 2 são apresentadas as técnicas utilizadas para extração de informação em textos e imagens. Na Seção 3 é apresentada a base de dados. O método para a recomendação de grupos e os algoritmos de aprendizado de máquina usados nos experimentos são descritos na Seção 4. Os resultados experimentais e a sua análise são realizados na Seção 5. Finalmente, a Seção 6 termina o trabalho com as conclusões e trabalhos futuros.

2. Técnicas de Extração de Informação

2.1. Representação de Textos

Quando se trabalha com textos é necessário realizar uma série de procedimentos para facilitar a recuperação da informação. A seguir são descritos os passos seguidos neste trabalho.

O primeiro passo a ser realizado é o pré-processamento do texto. Nesta etapa são removidos pontuação, caracteres não alfanuméricos e números. A seguir, as letras são passadas para caixa baixa e termos com baixo valor semântico, conhecidos como *stopwords* [Baeza-Yates e Ribeiro-Neto, 1998], são removidos, pois além de não contribuírem com informações relevantes, eles podem prejudicar a tarefa de recuperação de informação. Alguns exemplos de *stopwords* são artigos, preposições e interjeições.

Após este procedimento, é utilizada uma etapa para obter a forma canônica das palavras, isto é, sem as flexões verbais, diminutivos, aumentativos e plural. Esta etapa é importante pois o computador não é capaz de entender, sem este processamento, que palavras como, professor e professores estão se referindo a um mesmo tipo de profissão. Neste trabalho foi utilizado o software Cogroo (cogroo.sourceforge.net) que permite, entre outras coisas, obter a forma canônica de palavras da língua portuguesa.

Um próximo passo é adotar uma forma eficiente para representar um texto tal que possa ser processado por um computador de forma adequada. Uma das formas mais populares de representação de textos é através do modelo de espaço vetorial apresentado por Salton e Yang [1975]. Neste modelo, cada documento (ou texto) d_j é representado por um vetor de n dimensões, sendo que cada posição t_i deste vetor representa um termo que está associado a um peso $w_{i,j}$. Neste trabalho, $w_{i,j}$ é o número de vezes que o termo t_i ocorreu dentro do documento d_j. Dessa forma, o documento d_j é representado conforme a Equação 1:

$$d_j = \{w_{1,j}, w_{2,j}, \ldots, w_{n,j}\}. \tag{1}$$

Para um conjunto de D documentos, é obtido um conjunto de termos distintos que formam o dicionário de termos de D. Quando um termo não ocorre dentro de um documento é associado um peso zero a ele. Com a representação vetorial é possível utilizar métricas como distância euclidiana e similaridade de cosseno para medir o grau de semelhanças entre diferentes documentos. Na tarefa realizada neste trabalho, cada documento está associado a uma categoria presente no conjunto C, que são os grupos do Facebook.

Dois passos adicionais são realizados sobre o conjunto de documentos D. O primeiro é a remoção de termos poucos discriminativos, que neste trabalho são considerados os termos que apareceram em apenas um documento de D ou que apareceram em todos os documentos de D. No primeiro caso, a remoção é motivada pois pode ser um erro de ortografia ou palavra muito rara que não influencia nos resultados, mas que aumenta o número de termos presentes nos documentos, aumentando o custo computacional na etapa de classificação. No segundo caso, a presença de termos em todos documentos pode indicar ser uma *stopword* que não foi removida dos textos.

O segundo passo adicional corresponde ao uso da técnica de ponderação dos pesos dos termos, conhecida como *Inverse Document Frequency* (idf), que é utilizada para aumentar a importância dos termos que aparecem em poucos documentos e diminuir daqueles termos que aparecem em muitos documentos [Salton e Yang, 1975]. Essa ponderação é combinada com a frequência de ocorrência de cada termo em cada documento, sendo obtido o novo peso $\widehat{w_{i,j}}$ do termo t_i no documento d_j. Este novo peso é calculado pela Equação 2:

$$\widehat{w_{i,j}} = w_{i,j} \times idf_i, \quad idf_i = \log\left(\frac{|D|}{nd_i}\right), \tag{2}$$

onde $w_{i,j}$ é a frequência do termo t_i no documento d_j, idf_i é o valor do idf para o termo t_i, sendo $|D|$ o número de documentos do conjunto D e nd_i é o número de documentos que o termo t_i apareceu, ou seja, o número de documentos em que $w_{i,j} > 0$.

Assim, ao final de todos os procedimentos, é obtida uma matriz numérica M com dimensões $|D| \times n$, representando os documentos, e um vetor N com dimensões $|D| \times 1$, representando o vetor de categorias associados aos documentos, sendo n o número de termos distintos no conjunto D.

2.2. Representação de Imagens

Assim como os textos, as imagens também precisam passar por uma etapa de pré-processamento para extrair informações relevantes e reduzir a dimensionalidade dos dados.

Existem várias formas propostas para representar imagens em tarefas de categorização, porém muitas delas são lentas, complexas, necessitam de calibrar muitos parâmetros ou precisam de um conjunto de imagens com vários elementos rotulados [de Souza Gazolli, 2014]. Neste trabalho foi adotado a abordagem proposta por [de Souza Gazolli e Salles, 2012], por ser rápida, simples de implementar, sem parâmetros para calibração e apresentar resultados semelhantes aos de outras abordagens.

A abordagem usada neste trabalho, chamada de Transformada Census Modificada Contextual (CMCT), representa cada imagem na forma de um vetor. Nesta abordagem, inicialmente é obtida a imagem em escala de cinza. Sobre cada pixel x da imagem é centralizada uma pequena máscara 3×3 que realiza uma operação sobre os pixels vizinhos ao pixel x. Primeiro é calculado a intensidade média $\overline{I_x}$ dos pixels embaixo da máscara. A seguir, as intensidades dos pixels vizinhos ao pixel x (N_x) são comparadas a $\overline{I_x}$. Se o valor da intensidade de um pixel y (I_y) for maior ou igual a $\overline{I_x}$, um bit 1 é gerado no pixel da máscara 3×3 que está sobre o pixel y. Caso contrário, é gerado um valor 0 no mesmo local da máscara. A Equação 3 apresenta este cálculo. Após passar por todos os pixels da vizinhança de x, a máscara 3×3 estará com valores binários em todas as posições, exceto no pixel central. Esses valores binários formam uma palavra de 8 bits que é convertida em um número inteiro entre 0 e 255. Esta operação inicial é chamada de Transformada Census Modificada (MCT), ou MCT8 por formar palavras de 8 bits.

$$T_x = \otimes_{y \in N_x} \zeta(I_y, \overline{I_x}), \quad \zeta(m,n) = \begin{cases} 1, & m \geq n \\ 0, & \text{caso contrário.} \end{cases} \tag{3}$$

O número obtido na Equação 3 é armazenado numa posição relacionada ao pixel x para gerar uma nova imagem. Ou seja, após passar a máscara em todos os pixels da imagem original, uma nova imagem é obtida. Dessa nova imagem é calculado o seu histograma, ou seja, o número de vezes que cada valor ocorreu na nova imagem. Este histograma possui $2^8 = 256$ elementos.

Após este passo, o MCT é passado sobre essa nova imagem e obtido o histograma da imagem resultante, novamente com 256 elementos. Os histogramas da primeira e da segunda imagem gerada são concatenados formando um vetor h de 512 elementos. Para evitar uma diferença muito grande entre os valores dos elementos, e obter um melhor desempenho de classificação, é usada uma operação de logaritmo sobre os valores não nulos do vetor h, conforme indicado na Equação 4. Após este passo, o novo vetor $\tilde{h}$ é normalizado pela Equação 5.

$$\tilde{h}_i = 1 + \log(h_i) \quad \forall\, i \in \{1, \ldots, 512\} | h_i > 0. \tag{4}$$

$$\widehat{h_i} = \frac{\tilde{h}_i}{\sum_{i=1}^{512} \tilde{h}_i} \quad \forall\, i \in \{1, \ldots, 512\}. \tag{5}$$

O procedimento anterior é aplicado para cada imagem da base de dados. Ao final, é obtida uma matriz numérica F com dimensões $W \times 512$ representando as imagens e um vetor N com dimensões $W \times 1$, representando o vetor de categorias associados às imagens, sendo W o número de imagens na base de dados.

3. Base de Dados

Para a realização dos experimentos foram coletados posts de seis grupos da rede social Facebook, no período entre 01/04/2011 a 15/07/2015, usando a API em java Facebook4j (facebook4j.org). Os grupos são sobre diferentes assuntos, como natureza, religião, política, etc. De cada post foi coletado um conjunto de campos, como mensagem, data de postagem, identificação do post, nome do grupo, link da imagem postada no Facebook, entre outros.

Uma vez coletados os posts, o segundo passo foi a coleta das imagens. Nesta etapa, somente as imagens presentes no Facebook foram coletadas, visto que links que apontavam para páginas externas poderiam estar quebrados ou terem sido atualizados, tal que poderiam mostrar imagens diferentes das originais postadas. Além disso, só foram coletadas imagens que tivessem pelo menos uma dimensão mínima de 15×15 pixels e tamanho mínimo de 15 kB.

Após a coleta das imagens, um procedimento de filtragem sobre os dados coletados foi aplicado. Este procedimento teve como função remover todos os posts cujo ao menos um campo, seja de mensagem ou de link de imagem, estivesse vazio. Também é considerado um link de imagem vazio se houver um link, mas não foi coletada a imagem. Finalmente, após estas etapas, obteve-se a base de dados usada neste trabalho.

3.1. Preparação dos dados

Para realizar a tarefa de recomendar grupos aos usuários de acordo com os seus perfis de postagens, os seguintes procedimentos foram realizados. Inicialmente, foram identificados os 20 usuários com o maior número de postagens na base de dados. Todos os posts desses 20 usuários foram separados da base de dados, e com eles serão obtidos informações dos perfis de postagens desses usuários para poder recomendar grupos a eles. Estes dados serão chamados de base de usuários.

Removidos todos os posts desses 20 usuários da base de dados, o passo seguinte foi selecionar aleatoriamente, de cada grupo, 1000 posts. Esse número foi selecionado para que o número de posts por grupo fosse igual e evitar que um grupo tenha mais dados que os outros. Esses posts serão usados para obter informações de cada grupo. Esses dados serão chamados de base de grupos.

Para cada post da base de usuários e de grupos foi obtida a representação em forma de vetor da mensagem, conforme o procedimento descrito na Seção 2.1. Assim, foi obtida uma matriz de mensagens da base de usuários com 4830×6183, ou seja, 4830 posts e 6183 termos diferentes, enquanto que as mensagens para a base de grupos formaram uma matriz de 6000×6183. Esse procedimento foi seguido tomando o cuidado de garantir que os índices de palavras em ambas as matrizes de dados fosse o mesmo. Por exemplo, se na primeira matriz a palavra "descansar" estivesse na posição 45, na segunda matriz também seria 45.

Também foi obtida para cada base de dados a representação vetorial de cada imagem incluída nos posts. Para isso, o procedimento apresentado na Seção 2.2 foi utilizado. Para a base de dados de usuários, foi obtida uma matriz de imagens de 4830×512 e, para a base de grupos, uma matriz de 6000×512.

Assim, cada post de cada base de dados é representado por dois vetores: um vetor representando a mensagem e outro vetor representando a imagem. Com essas matrizes serão realizados os experimentos, conforme descrito na Seção 4.

4. Método de Recomendação de Grupos

A abordagem proposta para recomendar os grupos aos usuários se baseia no perfil dos posts de cada usuário. Um algoritmo de classificação será inicialmente treinado com uma das matrizes obtidas da base de grupos, que pode ser a de mensagens ou a de imagens. Após o treinamento, será apresentada ao classificador a matriz equivalente da base de usuários. Ou seja, se o classificador foi treinado com uma matriz de mensagens, a matriz apresentada ao classificador será também uma matriz de mensagens.

Sendo assim, o classificador retornará o grupo (classe) mais provável a ser associado ao vetor avaliado. Ao final do processo, todos os vetores da matriz de usuários estarão associados a

um grupo pelo classificador. Após esta etapa, é identificado para cada usuário o número de posts associado a cada um dos grupos, de tal forma que será obtido um vetor com g dimensões, sendo g o número de grupos, que neste trabalho é igual a 6. A partir desse vetor são recomendados os grupos para o usuário. A Tabela 1 mostra um exemplo fictício desse resultado para dois usuários e três grupos.

Tabela 1: Exemplo da categorização de posts de 2 usuários.

Usuários	**G1**	**G2**	**G3**
usuário_1	23	6	10
usuário_2	18	5	33

Uma abordagem simples para a recomendação seria recomendar os grupos cujo o usuário tenha ao menos um post rotulado neste grupo. Porém, como pode ser visto na Tabela 1, é comum o usuário ter posts rotulados em todos os grupos, e essa abordagem iria falhar ao recomendar todos os grupos. Uma alternativa proposta é aplicar um limiar τ sobre os números de posts. Todos os grupos que possuem um número de posts maior ou igual a τmax_{post} é recomendado ao usuário, onde max_{post} é o número de posts do grupo com o maior número de posts. Para o caso da Tabela 1, se $\tau = 0,5$, o usuário_1 é associado à G1 e o usuário_2 à G1 e G3.

Existem vários classificadores que poderiam ser usados neste trabalho, sendo neste trabalho usado dois: o kNN (k Vizinhos mais Próximos) e SVM *Support Vector Machine* (em ambos os casos foram usados implementações de terceiros). O kNN foi usado por questão de simplicidade de implementação e entendimento. O kNN é um classificador que calcula a distância entre cada vetor de treinamento e o vetor que será classificado. A partir deste resultado, são selecionados os k vizinhos mais próximos do vetor de teste, que são os k vetores de treinamento mais próximos do vetor de teste. A classe (grupo) com mais vetores entre os k vizinhos mais próximos é associada ao vetor de teste Duda et al. [2001]. A medida mais comum a ser aplicada no kNN para textos é a similaridade cosseno [Baeza-Yates e Ribeiro-Neto, 1998], sendo portanto a utilizada neste trabalho. A Equação 6 mostra esse cálculo, onde x e y são dois vetores e d é o número de elementos nos vetores.

$$cos(x,y) = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} {x_i}^2}\sqrt{\sum_{i=1}^{d} {y_i}^2}}. \tag{6}$$

O SVM é mais complexo, mas foi usado porque obteve melhores resultados que o kNN para classificação de imagens, segundo o observado em [de Souza Gazolli, 2014]. O SVM encontra hiperplanos que conseguem separar, com uma margem de separação maximizada, exemplos (vetores) de diferentes classes (grupos). Caso os exemplos não sejam linearmente separáveis, é necessário usar um fator de custo c para definir o efeito da separação errada dos exemplos. Além disso, pode-se mapear os exemplos em um novo espaço para facilitar a separação, geralmente um espaço de maior dimensão. Para isso é usado um método chamado de kernel [Bennett e Campbell, 2000]. Neste trabalho foi utilizado o kernel de interseção de histograma [Barla et al., 2003], o mesmo utilizado por [de Souza Gazolli, 2014]. Para o SVM utilizado é necessário definir o valor de c.

Para avaliar a qualidade dos resultados do método proposto foi utilizada a métrica acurácia, que consiste no número de usuários cujos grupos foram recomendados corretamente dividido pelo número de usuários, multiplicada por 100%. Esse valor varia entre 0 a 100%, e quanto maior for, melhor.

5. Experimentos e Resultados

Antes de usar a abordagem proposta, é necessário definir o limiar τ, e os valores de k e c do kNN e SVM, respectivamente. Foram testados um conjunto de valores em uma etapa de validação de forma a maximizar a acurácia do método. Para isso, foram testados valores de k na

Tabela 2: Resultados dos experimentos com os dados separados.

Tipo de dados	kNN	SVM
Mensagens	**73,33%**	66,67%
Imagens	53,33%	**66,67%**

Tabela 3: Resultados dos experimentos com os dados juntos.

Método	kNN	SVM
Dois classificadores	**73,33%**	66,67%
Mensagens + imagens	73,33%	**80,00%**

faixa entre 1,2,...,15 e limiares de τ entre 0,4 e 1,0, aumentando em passos de 0,1. Os diferentes valores de k e τ foram avaliados obtendo a acurácia sobre 5 dos 20 usuários e usando os 6000 posts dos grupos para treinar o kNN. Os valores de k e τ que retornaram a maior acurácia foram usados para recomendar os grupos para os demais 15 usuários. Foram encontrados valores de k e τ para os dados de mensagens e de imagens. Procedimento semelhante foi usado para o SVM: os mesmos 5 usuários foram usados para obter os melhores valores de c e τ, tanto para os dados de mensagens como de imagens, sendo o SVM treinando com os dados dos grupos. Os melhores valores foram usados para recomendar os grupos para os demais 15 usuários. Os valores de c avaliados foram entre 1 e 15, aumentando em passos de 2.

A segunda linha da Tabela 2 mostra a acurácia obtida sobre os 15 usuários quando usando somente a informação das mensagens, enquanto que a terceira linha mostra os resultados quando usando somente as imagens como informação. A segunda coluna mostra os resultados obtidos pelo kNN e a terceira mostra os resultados do SVM.

Como pode ser observado, o kNN obteve um resultado melhor do que o SVM para recomendar grupos a partir das mensagens, enquanto o SVM foi melhor que o kNN quando é usada a informação proveniente das imagens. Também é observado que as mensagens foram mais informativas que as imagens, pois as acurácias obtidas com as mensagens foram maiores ou iguais às obtidas pelas imagens. No entanto, as imagens mostraram ser uma fonte de informação útil, pois só com esse tipo de informação foi possível recomendar corretamente os grupos, no pior dos casos, para 53,33% dos usuários, ou o equivalente a 8 usuários.

Um novo conjunto de experimentos foram realizados combinando as informações de texto e imagens, sendo essa combinação realizada de duas formas. Na primeira são treinados dois kNN, um com os dados de mensagens e outro com os dados de imagens, com os valores de k encontrados na etapa anterior. Assim, cada post é classificado duas vezes: o texto e a imagem, e é somado o número de posts associados a cada grupo. Novamente é encontrado um valor de τ sobre os 5 usuários para ser aplicado sobre os outros 15 usuários. O mesmo procedimento é usado para o SVM.

Na outra forma, os vetores de mensagens e imagens de cada post são concatenados, formando um único vetor por post. Os parâmetros de kNN e SVM são selecionados conforme explicado anteriormente, e são obtidos os resultados sobre os 15 usuários.

A Tabela 3 indica os resultados obtidos, sendo a segunda linha o resultado quando se usa dois classificadores e a terceira linha quando os vetores são concatenados. Os resultados do kNN e do SVM são mostrados na segunda e terceira coluna, respectivamente.

Como visto na tabela, os resultados de ambos os métodos foram iguais ou superiores ao uso de somente um tipo de informação, e para o caso da concatenação dos vetores, o SVM obteve uma melhora em relação ao uso dos dados de forma separada. Este resultado realça a importância de usar ambos os tipo de informação em tarefas de recomendação e classificação.

6. Conclusão

Neste trabalho foi analisado o uso de informações extraídas de textos e imagens divulgados pelos usuários para recomendar grupos do Facebook. Os resultados experimentais obtidos indicam que tanto textos como imagens carregam informações relevantes para realizar esta tarefa, e que a combinação desses dois tipos de dados pode fornecer melhores resultados.

Trabalhos futuros incluem o estudo de métodos para não indicar nenhum grupo, caso o usuário de fato não pertença e não tenha afinidade a nenhum dos grupos analisados. Também será estudado formas de usar outros campos coletados na base de dados original, como por exemplo o campo tempo, com intuito de aprimorar os resultados. Por fim, o aumento do número de usuários avaliados permitirá uma análise mais precisa dos resultados.

7. Agradecimentos

Os autores gostariam de agradecer à Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) pelo apoio financeiro, com o projeto nº 526/2015.

Referências

Alqhtani, S. M., Luo, S., e Regan, B. (2015). Fusing text and image for event detection in twitter. *The International Journal of Multimedia & Its Applications (IJMA)*, 7(1):27 – 35.

Baeza-Yates, R. e Ribeiro-Neto, B. (1998). *Modern Information Retrieval*. Addison-Wesley, New York, first edition edition.

Barla, A., Odone, F., e Verri, A. (2003). Histogram intersection kernel for image classification. *Proceedings of 2003 International Conference on Image Processing*, 3:513–516.

Bennett, K. P. e Campbell, C. (2000). Support vector machines: Hype or hallelujah? *SIGKDD Explorations Newsl.*, 2(2):1–13.

de Souza Gazolli, K. A. (2014). *Utilizando Contexto na Representação de Imagens para a Classificação de Cenas*. PhD thesis, Universidade Federal do Espírito Santo.

de Souza Gazolli, K. A. e Salles, E. O. T. (2012). A contextual image descriptor for scene classification. *Online Proceedings on Trends in Innovative Computing*, p. 66–71.

Duda, R. O., Hart, P. E., e Stork, D. G. (2001). *Pattern Classification*. Wiley-Interscience, New York, 2 edition.

Jin, L., Chen, Y., Wang, T., Hui, P., e Vasilakos, A. V. (2013). Understanding user behavior in online social networks: a survey. *IEEE Communications Magazine*, p. 144–150.

Kosinski, M., Stillwell, D., e Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, p. 1–4.

Mislove, A., Viswanath, B., Gummadi, K. P., e Druschel, P. (2010). You are who you know: inferring user profiles in online social networks. *Proceedings of the third ACM international conference on Web search and data mining*, p. 251–260.

Salton, G. e Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, p. 613–620.

Statista (2016). Number of social network users worldwide from 2010 to 2019 (in billions). Technical report, http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

WERSM (2015). How much data is generated every minute on social media? Technical report, http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/.