Aprendizagem em Redes Socias: uma Análise de Dados do Twitter

Guilherme M. Torres
Univ. Federal De São Carlos Rdv João Leme Dos Santos, Km 110 – CEP 18052-780 – Sorocaba – SP - Brasil
guimato@gmail.com

Luciana A. M. Zaina
Univ. Federal De São Carlos Rdv João Leme Dos Santos, Km 110 – CEP 18052-780 – Sorocaba – SP - Brasil
lzaina@ufscar.br

Tiago A. de Almeida
Univ. Federal De São Carlos Rdv João Leme Dos Santos, Km 110 – CEP 18052-780 – Sorocaba – SP - Brasil
talmeida@ufscar.br

Resumo

A Web 2.0 fez com que o uso de aplicações relacionadas a redes sociais tenham crescido. O *Twitter* tem se destacado por ser um meio de colaboração, comunicação e de troca de ideias de pessoas com interesses em comum. Este artigo apresenta um algoritmo que tem como objetivo buscar padrões de intersse em mensagens do *Twitter*, a partir de um conjunto de palavras-chave em um ambiente de aprendizagem. Um experimento foi realizado durante o segundo semestre de 2011 com um grupo de alunos que seguiam docentes da área de Computação. Foi realizada uma análise das mensagens coletadas.

Categorias

H.1.2 [**User/Machine Systems**]: Human information processing

Termos Gerais

Experimentation

Palavras-chaves

Tokenização, minerção de dados, redes sociais, aprendizagem colaborativa.

1. INTRODUÇÃO

O surgimento da Web 2.0 fez com que aplicações relacionadas a redes sociais fossem amplamente utilizadas. Isto porque suas características favorecem a expressão e socialização por meio de ferramentas de comunicação e colaboração como *blogs*, *wikis* e redes sociais em geral [1]. As redes sociais online sociais surgiram com diferentes propósitos sendo alguns exemplos: rede de profissionais, *LinkedIn*, redes para compartilhamento de mensagens curtas, *Twitter*, compartilhamento de vídeos, *Youtube*, redes de amigos, *Facebook* [2].

As redes sociais tem fomentado a utilização de seus ambientes para a aprendizagem. Nelas os usuários têm a possibilidade de expressar seu conhecimento e compartilhar este com diferentes pessoas que possuem os mesmos interesses [3]. Dentre as diversas ferramentas de redes sociais o *Twitter* tem se destacado. O *Twitter* pode ser definido como um serviço de *microblogging*, que utiliza mensagens curtas (140 caracteres ou menos) para transmitir uma informação, podendo ser executado em diferentes dispositivos. Diariamente vinte milhões de usuários geram mais de cinquenta milhões de mensagens[4]. Analisar as mensagens postadas no *Twitter* pode auxiliar no levantamento de algumas informações sobre os usuários. Além de verificações simplesmente numéricas, como o número de seguidores que certa pessoa tem, é possível analisar as mensagens postadas através de técnicas de mineração de textos e análise de padrões. A mineração de textos tem sido utilizada na área de aprendizagem eletrônica com o objetivo de auxiliar não só o professor na identificação de sucessos e insucessos durante o processo de ensino-aprendizagem, como também para identificar, ou mesmo recomendar assuntos que possam contribuir com a aprendizagem do aluno [5]

O objetivo deste trabalho é propor um algoritmo com o objetivo de buscar padrões de interesse em mensagens do *Twitter*, a partir de um conjunto de palavras-chaves. Para isto foram coletadas e analisadas mensagens do *Twitter,* postadas por docentes que estejam sendo utilizadas com a finalidade de aprendizagem. Para validar o algoritmo foi realizada uma experiência, onde foi analisado se os alunos que seguiam os docentes reencaminhavam as mensagens ou mesmo, postavam novas mensagens sobre o mesmo tema. Um questionário foi aplicado aos alunos que participaram da experiência com o objetivo de analisar os resultados obtidos a partir do algoritmo.

O restante do artigo está organizado da seguinte forma: Seção 2 faz uma breve discussão dos fundamentos teóricos do trabalho. Seção 3 relata alguns trabalhos relacionados. Seção 4 é proposto o algoritmo para análise dos dados. Seção 5 é feita a experimentação e análise dos resultados. Seção 6 as considerações finais sobre o trabalho.

2. FUNDAMENTAÇÃO TEÓRICA E TECNOLÓGICA

Para que o trabalho pudesse ser desenvolvido alguns conceitos sobre mineração de textos foram estudados, conforme descrito a seguir.

A mineração de textos pode ser definida como a procura por padrões em um texto em linguagem natural através de um processo de análise do texto, buscando extrair informações deste texto para um propósito em particular. A mineração de textos é dividida em etapas: identificação do problema, pré-

Workshop de Trabalhos de Iniciação Científica'12, 15 a 18 de Outubro, 2012, São Paulo, São Paulo, Brazil.

processamento, extração de padrões (*data mining*), pós-processamento e utilização do conhecimento [5].

O pré-processamento é a etapa de preparação dos textos que irão ser utilizados na mineração. Um dos problemas na mineração de textos é que os dados não se encontram estruturados implicando em uma limitação na utilização dos algoritmos de aprendizagem de máquina. Uma das maneiras de se estruturar os dados é transformando-os em uma matriz de atributo-valor na qual a frequência das palavras, independentemente do seu contexto, é contado. Para produzir uma tabela com atributo-valor relevante ao problema, é realizada a tokenização, que examina um texto não estruturado e identifica suas características importantes separando o texto em *tokens* (palavras). O processo de tokenização pode ser definido como um processo de análise léxica, que analisa uma entrada de linhas de caracteres e gera uma sequência de símbolos. Durante a tokenização é necessário remover alguns caracteres indesejados, como sinais de pontuação, separação silábica, marcações especiais e números, os quais, isoladamente fornecem pouca informação [7]. Outro conceito estudado está relacionado as *stopwords, que* são palavras que devem ser desconsideradas, pois não agregam significado para o algoritmo de aprendizagem. Também foram estudadas expressões regulares, que são métodos de identificar um padrão em um texto [6].

Também foi necessário estudar a API do *Twitter*[1] para poder recuperar as mensagens postadas. A API é dividida em três partes, duas API REST que permite os desenvolvedores a acessar todos os dados do *Twitter* incluindo atualizações, status dos dados e informações de um usuário.

3. TRABALHOS RELACIONADOS

Alguns trabalhos relacionados ao uso de redes sociais como ambientes de aprendizagem foram estudados [9],[10],[11], [9], [10].

O *Twitter* foi introduzido como uma ferramenta para compartilhar ideias em um curso, em que os alunos acompanharam os docentes. Um dos resultados observados pelos autores foi que os alunos conseguiam escrever de maneira concisa as novas descobertas que tinham sobre um determinado tema, compartilhando essas ideias com outros colegas [9].

Experimentos com alunos chineses foram realizados em um curso de língua inglesa. Após os experimentos um questionário foi aplicado aos alunos onde 62% deles relataram que gostaram da experiência com o *Twitter*. Também foram feitas pesquisas com o *Twitter* para educação em saúde, utilizando como ferramenta de feedback para artigos científicos [10].

O aumento da comunicação entre os alunos e docentes através do uso do *Twitter* foi relatado em [11]. Docentes recebem mensagens diretas dos estudantes sobre o curso. Também foi analisada as mensagens trocadas entre eles no *Twitter* para poder identificar as questões que mais lhe interessam.

Um estudo comparando as mensagens postadas no *Twitter* durante um determinado período e os termos mais citados naquele período é apresentado em [9], relacionando estes com o perfil das pessoas que postaram as mensagens (idade, sexo, localização geográfica, etc). Para analisar as mensagens e identificar os perfis foi utilizado o algoritmo de *Kohonem* para auto-organização de mapas através do aplicativo Viscovery SOMine[2]. Os autores concluíram que os termos mais citados estão diretamente relacionados a alguma característica do perfil de quem postou a mensagem.

Outro estudo sobre a relevância do *Twitter* como ferramenta para disseminação de informações durante um desastre natural é discutido em [10]. Observou-se as mensagens postadas durante dois desastres naturais diferentes. Foram que coletadas mensagens do *Twitter* que possuiam as *tags* relevantes a localização geográfica do desastre natural. As mensagens coletadas foram analisadas através do e-Data Viewer[3], buscando observar a formação de grupos que caracterizassem uma categoria relevante ao fenômeno natural. Concluiu-se que as categorias estão relacionadas ao fenômeno natural ocorrido.

4. ALGORITMO PROPOSTO

A partir da fundamentação teórica e dos trabalhos relacionados este trabalho propõe um algoritmo com o objetivo de buscar padrões de interesse em mensagens do *Twitter*, a partir de um conjunto de palavras-chaves. Para isso, é utilizada uma etapa de tokenização das mensagens e depois são criadas matrizes que permitem verificar a ocorrência e frequência de palavras nas mensagens. Para execução do algoritmo deve-se considerar um conjunto de mensagens postadas no *Twitter*.

A partir das listas de *tokens* extraídas de cada mensagem avaliada, é criada uma matriz de ocorrência **D(|T|,|M|)**, onde **T** é a lista dos *token*s e **|T|** sua respectiva cardinalidade; **M** o conjunto das mensagens extraídas do *Twitter* e **|M|** a cardinalidade. Seja **m**, cada mensagem extraída do conjunto de mensagens, onde **m** pode ser descrito como uma matriz de *tokens*, **m(t1t2t3t4...tn)**, para cada mensagem é utilizado a função de tokenização, ***ftoken*(m)** que retorna uma lista de *tokens* extraídos da mensagem. Em (1) é apresentado o trecho do algoritmo que retrata o processo de criação das matriz **D**:

```
j <- 0
  for each m∈M
    ftoken(m)
      for each t_k ⊂ m
        i <- search(T, t_k)
          if i != 0                                   (1)
             D(i,j) <- 1
          else
             T  <- T∪t_k
             D(|T|+1,j) <- 1
  j++
```

[1] http://dev.twitter.com/doc.

[2] http://www.viscovery.net/somine/

[3] http://www.cs.colorado.edu/~starbird/e-dataviewer.html

A função ***ftoken*(m)**, descrita no algoritmo representado em (2.1), tem o objetivo de realizar a análise léxica da mensagem. Dois conjuntos de delimitadores são usados para identificar os *tokens*. Isto é necessário porque nas mensagens do *Twitter* é muito comum existirem *links* dentro das mensagens, e estes *links* podem ser relevantes a análise. O primeiro conjunto (**Dlin**) especifica apenas delimitadores mais gerais como quebra de linha, espaço em branco e tabulação. O segundo (**DlinLink**) utiliza delimitadores mais comuns na escrita de textos como pontos e vírgulas, dois pontos, travessão, entre outros. A função de tokenização é executada em duas etapas. Primeiro, é lido caractere por caractere da mensagem, armazenando-os em **a**. Durante a leitura são ignorados os delimitadores mais gerais. Quando é encontrado um caractere que não é um delimitador geral, se inicia a etapa de criação do *token*, onde são concatenado os caracteres lidos em **a,** até que um delimitador geral seja novamente encontrado.

for each m∈M

while a == Dlin

a++ **(2.1)**

while a != Dlin

t_k <- $t_k \cup$ a

a++

Após a etapa (2.1) o algoritmo de tokenização, verifica se o *token* extraído não é um *link*. Para esta verificação é empregada uma expressão regular. Caso o *token* considerado não seja um *link*, um outro processo de tokenização, apresentado em (2.2) é executado utilizando o segundo conjunto de delimitadores.

if t_k != Link

While a == DlinLink

a++

While a != DlinLink

t_{k2} = $t_{k2} \cup$ a

a++ **(2.2)**

T <- T $\cup$ t_{k2}

else

T <- T $\cup$ t_k

Após a matriz de ocorrência ser criada, deve-se construir uma matriz de frequência. Seja **F(|T|)**, a matriz de frequência, onde **T** é a matriz que contém os rótulos dos *tokens*. Cada linha da matriz de frequência **F** possui correspondência com a linha que possui o respectivo rótulo do *token* da matriz **T**. Para cada $t_k \in$ **T**, percorre-se todas as mensagens em **D**, somando a frequência que em que t_k aparece nas mensagens. A construção de **F** é representada por (3):

$$\sum_{j=0}^{|M|} \mathbf{D}(t_k,j), \qquad (3)$$

A partir da definição de uma lista de termos considerados relevantes para o domínio, **R**, e considerando a matriz de frequência **F**, é construída uma matriz de frequência relevante **Y(|R|)**. O algoritmo que reporta a construção de **Y** é apresentado em (4):

for each cada t_k∈F

i = search(t_k,R)

if i != 0 **(4)**

j <- search(t_k,T)

Y(i) <- F(j)

São construídas duas matrizes de termos relevantes, uma considerando o docente e outra considerando os alunos. Após a definição do matriz de frequência relevante **Y**, para alunos e docentes, é realizada a intersecção desses conjuntos para verificar com que frequência os alunos e docentes utilizam os mesmos termos (5).

$$\mathbf{Y}_{\mathbf{teacher}} \cap \mathbf{Y}_{\mathbf{student}} \qquad (5)$$

5. EXPERIMENTAÇÃO E ANÁLISE DOS RESULTADOS

Para validação do algoritmo utilizou-se MatLab[4] na sua implementação. A escolha deste ambiente de desenvolvimento foi devido a seu desempenho com matrizes e com um grande volume de dados. O Matlab é um ambiente de programação para desenvolvimento de algoritmos, análise de dados, visualização e cálculo numérico de alto desempenho, integrando cálculo com matrizes, processamento de sinais e construção de gráficos.

Para a avaliação do algoritmo proposto foi realizado um experimento, considerando as mensagens postadas no *Twitter* de dois docentes que atuam nas áreas de desenvolvimento para Web, engenharia de software e empreendedorismo; e dos alunos que seguiam esses docentes. Durante o segundo semestre de 2011 foram recuperadas, a cada quinze dias, durante cinco meses, mensagens postadas pelos docentes. Foram também recuperadas as mensagens de 52 alunos que seguiam esses docentes. Criando uma base de dados para ser utilizada na avaliação contendo 1794 mensagens, sendo 118 dos professores, e 1676 dos alunos. É importante destacar que os docentes não fizeram nenhuma recomendação aos alunos que os seguiam. Isto porque se desejava observar o comportamento destes alunos em uma rede social, sem um direcionamento por parte do docente.

Considerando a base de dados das mensagens coletadas, foram criadas as matrizes de frequência dos termos relevantes dos alunos e dos professores utilizando os algoritmos propostos na Seção 3. Os docentes indicaram os termos que deveriam ser considerados na matriz de termos relevantes, termos estes relacionados às disciplinas ministradas pelos docentes. Através da execução do algoritmo observou-se que os alunos que acompanhavam os docentes no *Twitter* realizavam com baixa frequência o reenvio de mensagens ou mesmo a postagem de mensagens que contivessem os termos relevantes. Dos 36 termos relevantes considerados pelos docentes constatou-se que apenas 2 termos foram mencionados pelos alunos em novas postagens ou

[4] http://www.mathworks.com/products/matlab/

reenvios. Para fundamentar as conclusões sobre o uso do *Twitter* com os alunos, neste experimento, foi elaborado um questionário com oito perguntas que buscavam identificar quais eram os termos que mais interessava os alunos, a frequência que eles usavam o *Twitter* e se as mensagens dos professores estavam contribuindo para a aprendizagem deles. O objetivo era verificar se os alunos eram agentes apenas receptores. Ou seja, os alunos acompanharam as mensagens postadas, mas não as repassava aos seus seguidores.

Dos 52 alunos que tiveram suas mensagens coletas, 38 responderam o questionário, ou seja, 73%. Alguns dados importantes foram levantados a partir das respostas dos alunos. Para a pergunta sobre a frequência com que os alunos repassavam as mensagens sobre os termos relevantes, praticamente 80% dos alunos responderam que a frequência com que eles repassam é inferior a dois.

Dentre os termos que os alunos relataram no questionário, os que mais se destacaram foram: "Internet das coisas", "Web", "Android", "Dispositivos móveis", "Google" e "*smathphones"*. Diferindo assim com os dados obtidos pela mineração das mensagens já que os alunos apenas reenviaram mensagens sobre "jobs" e "google". Cerca de 70% dos alunos relataram que acessam os links contidos nas mensagens; e 87% afirmaram que acessam os links relacionados aos termos considerados relevantes. Após ler alguma notícia sobre esses termos, 68% dos alunos relataram que buscam mais informações sobre esse tema na web. Porém, todos os alunos (100%) responderam que as mensagens que eles receberam contribuíram para adquirir novas informações.

Após a comparação dos resultados obtidos com a execução do algoritmo proposto e com a aplicação do questionário, conclui-se que os alunos são agentes receptores de informações no ambiente *Twitter*. Através desta conclusão observa-se a necessidade de que os ambientes de redes sociais necessitam de funcionalidades diferenciadas para a aprendizagem.

6. CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo propor um algoritmo para buscar padrões de interesse em mensagens do *Twitter*, a partir de um conjunto de palavras-chaves. Para validar o algoritmo foram coletadas e analisadas mensagens de docentes que estavam sendo utilizadas com a finalidade de aprendizagem. Um questionário foi aplicado aos alunos que participaram da experiência, buscando comparar os resultados obtidos com o algoritmo e a opinião direta dos alunos.

A partir dos resultados obtidos com a execução do algoritmo proposto e com a aplicação do questionário, conclui-se que os alunos são agentes receptores de informações no ambiente *Twitter*. Através desta conclusão observa-se a necessidade de que os ambientes de redes sociais necessitam de funcionalidades diferenciadas para a aprendizagem. Como futuro trabalho está sendo planejada uma comparação dos links existentes nas mensagens com os termos relevantes considerados na experimentação. O objetivo é tentar analisar se os alunos reenviam os links a outras pessoas através do *Twitter*.

7. AGRADECIMENTOS

Agradecemos a FAPESP pelo apoio financeiro.

8. REREFÊNCIAS

[1] Recuero, Raquel (2009). Redes Sociais Na Internet. Editora Sulina, 2009.

[2] Nielsen Online (2009). "Social networks & blogs now 4th most popular online activity, ahead of personal email". Disponível em: http://www.nielsen-online.com/pr/pr_090309.pdf. Acessado em: 20/01/2011.

[3] Dabbagh, Nada; Reo, Rick (2010). "Back to the future: tracing the roots and learning affordances of social software". In: Lee, Mark J.W; McLoughlin, Catherine, Web 2.0-Based E-Learning: Applying Social Informatics for Tertiary Teaching, 2010.

[4] Cormode, Graham; Krishnamurthy, Balachander; Willinger, Walter (2010). "A Manifesto for Modeling and Measurement in Social Media". First Monday (Online), Vol. 15, N. 9, 2010. Disponível em: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3072. Acessado em: 19/10/2010.

[5] Machado, Aydano; Ferreira, Rafael; Bittencourt, Ig Ibert; Elias, Endhe ; Brito, Patrick ; Costa, Evandro de Barros (2010). "Mineração de Texto em Redes Sociais aplicada à Educação a Distância". Colabor@ (Curitiba), Vol. 6, N. 23, p. 1-21, 2010.

[6] Dunlap, Joanna C. ; Lowenthal, Patrick R. (2009). Tweeting the night away: Using Twitter to enhance social presence. Journal of Information Systems Education, Vol. 20, N. 2, pp. 129-136, 2009.

[7] Borau, Kerstin; Ullrich, Carsten; Feng, Jinjin; Shen, Ruimin (2009). "Microblogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence". ICWL 2009, LNCS 5686, pp. 78–87, 2009.

[8] JSOnline (2009). "Professors experiment with Twitter as teaching tool". Disponível em: http://www.jsonline.com/news/education/43747152.html. Acessado em: 27/06/2012.

[9] Marc Cheong and Vincent Lee. 2009. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In Proceedings of the 2nd ACM workshop on Social web search and mining (SWSM '09).

[10] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microbloggingduring two natural hazards events: what twitter may contribute to situational awareness. In Proceedings of the 28th international conference on Human factors in computing systems (CHI '10).