Conhecimento em redes sociais: um estudo de caso aplicado ao *Twitter*

Miguel Airton Frantz, Angelo Augusto Frozza, Reginaldo Rubens da Silva

Instituto Federal Catarinense (IFC) – Campus Camboriú
Rua Joaquim Garcia S/N – 88.340-055 – Camboriú – SC

frantz.miguel@gmail.com, {frozza,reginaldo}@ifc-camboriu.edu.br

Resumo. *Este artigo descreve a criação de uma base de conhecimento a partir de dados não estruturados que um usuário produz. O foco consiste no desenvolvimento de uma aplicação que extrai o conhecimento dos textos publicados no Twitter. Como resultado, apresenta-se o conhecimento implícito presente na base de dados do usuário na forma de uma nuvem de tags contendo os termos relevantes que mais aparecerem na base de dados em questão.*

1. Introdução

Os avanços da tecnologia e a facilidade no compartilhamento de informações e conhecimentos, faz com que, diariamente, milhares de pessoas produzam informações novas que circulam pelo mundo. Essas informações compartilhadas criam imensos bancos de dados pessoais que, quando não estão organizadas ou classificados, acabam tornando difícil a realização de qualquer busca ou até mesmo seu entendimento.

Para Sousa (2012), o ato de classificar acompanha o cotidiano dos seres humanos, muitas vezes de forma imperceptível. Como exemplo, a rede social *Twitter* utiliza um sistema de classificação e indexação opcional das mensagens, que consiste em atribuir uma *tag* (etiqueta) à mensagem para classificá-la e representá-la. Como o *Twitter* não possui comunidades, fóruns ou grupos, as *tags* assumem o papel de realizar a classificação da informação nesta rede social. Esse sistema de classificação e indexação (manual), por muitas vezes acaba não sendo utilizado, o que faz com que as mensagens não possuam nenhuma classificação.

Assim sendo, surge a proposta de um mecanismo de Recuperação da Informação (RI) capaz de extrair as principais informações presentes nos textos publicados por um usuário em sua conta do *Twitter*. Utilizando técnicas de RI, é realizada a busca de palavras que possam expressar o sentido de cada *tweet*, ou seja, um conjunto de palavras chave. A junção destes termos, que podem ser considerados relevantes, ajuda a definir o perfil de determinado usuário e retrata os assuntos que ele mais aborda em sua conta. Para apresentar visualmente as informações coletadas, optou-se por utilizar nuvens de *tags*, nas quais as palavras são dispostas aleatoriamente, variando seu tamanho de acordo com o número de vezes que a mesma aparece no conjunto de texto analisado.

O restante deste artigo está organizado da seguinte maneira: na seção 2 são apresentados os materiais e métodos; a seção 3 apresenta os principais resultados obtidos; a seção 4 apresenta as considerações finais.

2. Materiais e Métodos

A aplicação proposta, denominada *TweetKnowledge*[1], tem sua arquitetura estruturada em camadas, de forma que cada camada procura ser, tanto quanto possível, independente das demais. Optou-se pelo desenvolvimento em camadas para que, em projetos futuros, seja possível reaproveitar facilmente os componentes de *software* produzidos:

a) *Camada de autenticação*: é responsável pelo processo de identificação do usuário. Para isso, utiliza-se o *login* social do *Twitter*, que autoriza o acesso aos dados cadastrais do usuário. Feita a autenticação, a aplicação recebe algumas informações sobre o usuário, que são utilizadas para realizar o controle de quem está submetendo uma base de dados, além de permitir a publicação da nuvem de *tags* no *Twitter*;
b) *Camada de upload dos arquivos*: é responsável por armazenar (*upload*) no banco de dados da aplicação os arquivos de *tweets* do usuário, os quais estão em formato JSON. O *upload* do histórico de *tweets* do usuário dá um ganho de performance considerável em comparação com a opção de buscar os dados de forma *on-line*;
c) *Camada de manipulação dos dados e indexação das palavras*: após o *upload* de um arquivo é feita a indexação do texto. Esta camada teve como base a infraestrutura de *Full Text Searching* (FTS) fornecida pelo próprio *PostgreSQL* (através do módulo *TSearch* 2). O FTS consiste em uma técnica de pesquisa e recuperação de informações de texto armazenado no banco de dados. Foi necessário compreender como funciona a indexação e os dicionários de palavras, uma vez que os mesmos são utilizados em diversas etapas de filtragem e limpeza dos dados;
d) *Camada de apresentação e geração da nuvem de tags*: responsável pela geração da nuvem de *tags* através do *framework D3-Cloud* (*https://github.com/jasondavies/d3-cloud*). Também permite compartilhar a nuvem de *tags* do usuário.

A dificuldade em classificar e encontrar informações não é algo presenciado somente no *Twitter*. Entretanto, escolheu-se utilizá-lo por se tratar de uma das três maiores redes sociais ativas, com aproximadamente 500 milhões de usuários (MASHABLE, 2012). Redes sociais permitem que usuários criem conteúdo e vêm se tornando chave em pesquisas relacionadas ao tratamento de grandes quantidades de dados, além de constituírem um ambiente propício para extração de conhecimento (BENEVENUTO, ALMEIDA e SILVA, 2012). Com o intuito de facilitar a interpretação das informações obtidas pela aplicação, optou-se por apresentar o resultado final no formato de uma nuvem de *tags*. Além disso, este formato permitirá fazer a ligação dos termos com outras bases de dados, de uma maneira muito natural para o usuário final.

3. Resultados e Discussão

O *TweetKnowledge* consiste em uma aplicação prática para demonstrar a resolução de problemas associados com Recuperação da Informação (RI) sobre uma base de dados produzida por um usuário que compartilha notícias e informações por meio de sua conta do *Twitter*. A RI sobre textos, em geral, consiste em diminuir o número palavras de um texto, sempre tomando o cuidado para que este não perca o seu sentido inicial. Assim, os textos são reduzidos até que reste o menor número possível de termos que expressem sentido. Esses termos podem ser entendidos como as palavras chave do texto em questão e são utilizados na aplicação desenvolvida neste projeto.

[1] *http://www.tweetknowledge.com*

Para realizar esta tarefa, o *PostgreSQL* trabalha com a indexação de textos e a análise dos documentos em *tokens*, que podem ser entendidos como fragmentos brutos do texto do documento. Também utiliza o termo "lexema" para referir-se às palavras consideras úteis para indexação e busca. Estes lexemas são obtidos após o tratamento dos *tokens* mencionados anteriormente. Dicionários são utilizados para realizar esta etapa. No *PostgreSQL* há vários dicionários padrões, mas também pode-se personalizar dicionários para necessidades específicas. O uso de dicionários permite realizar um controle refinado sobre como os termos (*tokens*) são normalizados, por exemplo: definir *stop words* - conjunto de palavras comuns que não possuem um sentido sozinhas, como artigos, preposições etc.; mapear sinônimos para uma única palavra - utilizando *ispell* (dicionário com palavras do idioma escolhido); mapear frases para uma única palavra - utilizando *thesaurus* (dicionário de sinónimos); mapear diferentes variações de uma palavra para uma forma canônica - utilizando regras de *Snowball stemmer* (redução das palavras por meio de algoritmos específicos).

Neste trabalho, as etapas descritas anteriormente resultam em um conjunto de lexemas presentes em cada *tweet*. Para obter a frequência que cada termo aparece, basta realizar uma contagem por meio de uma consulta SQL. Em função de ter-se adotado o paradigma da codificação por camadas (Figura 1), no caso da adaptação da aplicação para usar outras fontes de dados, é necessário reescrever apenas a camada de entrada de dados.

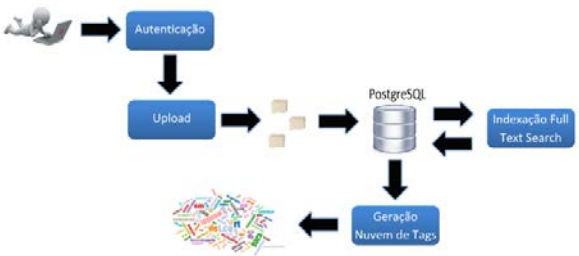

Figura 1 – Arquitetura da aplicação *TweetKnowledge*.

Figura 2 – Nuvem de *tags* gerada.

Percebe-se a importância desta pesquisa para a comunidade em geral, em especial, para aquelas pessoas que produzem algum tipo de conhecimento. Com isso, além de servir como mecanismo de indexação, o sistema permite que o usuário dê maior visibilidade à sua produção, por meio da nuvem de *tags* que pode ser compartilhada nas redes sociais.

Os termos apresentados na nuvem de *tags* (Figura 2) são as palavras consideradas relevantes após a manipulação e indexação dos dados na terceira camada. A quarta camada recebe como entrada uma lista de palavras e a frequência que cada uma aparece na base de dados e se encarrega de distribuí-las, determinando maior tamanho de fonte para palavras que tem maior número de repetições, além da variação das cores que ocorre de forma aleatória, com o intuito de facilitar a distinção entre os termos.

4. Considerações finais

Tendo em vista a importância da RI em redes sociais e, também, a relevância da classificação do conhecimento utilizando *tags* e palavras chave que representem uma mensagem ou texto, o presente artigo apresenta um mecanismo para extração de

conhecimento em cima das mensagens (*tweets*) de um usuário e permite a representação das informações obtidas de maneira que o usuário possa entender com facilidade.

Para tanto, é disponibilizada uma ferramenta que realiza a busca de palavras chave presentes em todo o histórico de publicações no *Twitter* de um usuário. Para obter os termos relevantes é realizado um tratamento dos textos anexados, utilizando dicionários de palavras, removendo as palavras que não possuem sentido e realizando diversas etapas de filtragem e limpeza dos dados. Após o tratamento dos textos, tem-se como resultado uma lista de termos relevantes, acompanhados do respectivo número de ocorrências. Com o objetivo de facilitar o entendimento e a assimilação da informação obtida, os termos relevantes são apresentados por meio de uma nuvem de *tags*. Um protótipo funcional do *TweetKnowledge* pode ser acessado através do *link http://www.tweetknowledge.com*.

Atualmente, a aplicação usa como entrada apenas os textos de *tweets*, porém, trabalhos futuros preveem o uso de outras fontes de conteúdo geradas pelos usuários, como outras redes sociais, documentos de textos (PDF, DOC etc.) e textos de *e-mails*. Além disso, pretende-se efetuar diversas melhorias, como utilizar *Linked Data* para adicionar conhecimento à lista de termos, sendo que cada termo da nuvem de *tags* direcionará para os textos (*tweets*) de sua origem e também para fontes externas.

Até o presente momento, foi encontrado apenas um trabalho que se propõem a fazer algo semelhante, chamado *TweetCloud* (*http://tweetcloud.icodeforlove.com*). Entretanto, segundo análise realizada, o mesmo apresenta diversos problemas, como lentidão por buscar os textos dos *tweets on-line* (o que também restringe a quantidade de texto a ser indexada), falhas na filtragem dos termos, utilização de um *layout* de nuvem simples, entre outros.

Referências bibliográficas

BENEVENUTO, F.; ALMEIDA, J.; SILVA, A. Coleta e análise de grandes bases de dados de redes sociais online. In: JORNADA DE ATUALIZAÇÕES EM INFORMÁTICA (JAI). Cap. 2. **Anais do XXXII Congresso da Sociedade Brasileira de Computação (CSBC)**. Curitiba: SBC, 2012.

FELIX, V. **Faça o download do histórico do Twitter**. In: <http://blogs.estadao.com.br/link/faca-o-download-do-historico-do-twitter/>. São Paulo: Estadão, 18 jan. 2013.

MASHABLE. **Will You Be Twitter's 500 Millionth User?** 2012. Disponível em: <http://mashable.com/2012/02/22/twitters-500-million-user/>. Acesso em: 25 jul. 2014.

POSTGRESQL. Chapter 12. **Full Text Search**. Disponível em: <http://www.postgresql.org/ docs/9.3/static/textsearch.html>. Acesso em: 20 mar. 2016.

SOUSA, A. M. de. **Organização em sistema caótico**: uso das tags para classificação da informação pelos usuários da rede social Twitter. 106 f. 2012. Mestrado (Mestrado em Ciência da Informação) – UFRJ, Rio de Janeiro (RJ).