TWITKNOWLEDGE - OBTENDO CONHECIMENTO A PARTIR DOS *TWITTES* PESSOAIS

Miguel Airton Frantz[1]*; Angelo Augusto Frozza*[2]

RESUMO

A proposta básica desse projeto consiste em criar uma base de conhecimento a partir de bancos de dados não estruturados que um usuário produz durante sua vida. O foco inicial está em desenvolver uma aplicação de Recuperação da Informação (RI) sobre os textos publicados na rede social *Twitter*, a qual representa apenas um dos formatos de entrada de dados possíveis para a aplicação. Como resultado, apresenta-se o conhecimento implícito presente na base de dados de *tweets* fornecida pelo usuário na forma de uma nuvem de *tags*, na qual estão os termos relevantes que mais aparecerem na base de dados usada.
Palavras-chave: Conhecimento. Dados. *Twitter*. Nuvem. *Tags*. Mineração.

INTRODUÇÃO

Com o desenvolvimento e a popularização da informática, um grande número de pessoas passou a ter acesso aos computadores e às facilidades associadas ao uso dos mesmos. Destaca-se, nesse sentido, a Internet, que se tornou um importante meio de comunicação e disseminação de conhecimento, oferecendo uma infinidade de recursos e passando a alcançar grande parte da população. Hoje, a Internet está presente em grande parte das casas, escolas, faculdades, empresas e em diversos outros locais, possibilitando assim, acesso às informações e notícias do mundo em praticamente todos os lugares.

É impressionante a facilidade e a rapidez com que hoje se pode encontrar informações na rede, sobre qualquer assunto e a qualquer hora. Essa facilidade está presente também no compartilhamento de informações, o que motiva muitas pessoas a publicar suas produções e compartilhar seus conhecimentos, tornando-os disponíveis na rede, seja através de *blogs*, redes sociais ou outros meios. Esse grande número de pessoas produzindo conteúdos e informações, e tornando esse material público, faz com que se tenha um enorme banco de dados.

É imensa a quantidade de conteúdo produzido e disponível hoje na Internet. Muitas vezes, é tanto conteúdo que as pessoas acabam se perdendo na infinidade de informações. Imensos bancos de dados pessoais fazem com que fique difícil fazer a recuperação dessas informações, uma vez que as mesmas não são estruturadas, ou seja, estão soltas dentro da organização/computador, sem ter sido tratadas e classificadas anteriormente. Informações como a que se encontram em arquivos de texto, planilhas eletrônicas, postagens em *blogs* e redes sociais, *e-mails*, entre outros, e que deixam os usuários limitados a buscas textuais oferecidas por mecanismos de busca como o *Google*. Esse problema fica ainda maior quando o usuário, produtor de conhecimento, começa a ter dificuldades de encontrar materiais que ele mesmo produziu sobre determinado assunto há algum tempo.

Nesse sentido, propõe-se um mecanismo de Recuperação da Informação baseado em nuvem de *tags* para uso das pessoas que produzem conhecimento. A partir

[1]Estudante de Graduação em Sistemas de Informação, IFC-Camboriú. Bolsista do CNPq Brasil. *E-mail*: frantz.miguel@gmail.com.
[2]Msc. em Ciência da Computação, UFSC; Professor do IFC-Camboriú. *E-mail*: frozza@ifc-camboriu.edu.br.

desse mecanismo, o usuário pode indexar seus textos e gerar informação em um formato de representação visual, através de uma nuvem de *tags*.

As nuvens de *tags* correspondem a um processo de indexação que leva em conta o número de vezes que determinado termo relevante aparece. Quanto maior o número de vezes que este termo aparece, maior é seu tamanho na representação da nuvem de *tags*, dando assim, maior destaque aos assuntos (termos) que são mais frequentes.

Dada a amplitude de possibilidades de formatos de entrada para esse mecanismo, esse projeto limitou-se aos textos publicados por um usuário no *Twitter*. Entre as principais redes sociais que podem ser aproveitadas para a descoberta de conhecimento e Recuperação da Informação está o *Twitter* (TEIXEIRA e DUQUE, 2012). Segundo estudo publicado pela agência Mashable (2012), esta é uma das três maiores redes sociais ativas, com aproximadamente 500 milhões de usuários.

As redes sociais na Internet tornaram-se ambientes ideais para estudo de diversas áreas, incluindo a Recuperação de Informação e a Descoberta de Conhecimento. Segundo Benevenuto (2012), as redes sociais permitem que usuários criem conteúdo e vêm se tornando um tema chave em pesquisas relacionadas à organização e tratamento de grandes quantidades de dados, além de constituírem um ambiente ideal para extração de conhecimento e aplicação de técnicas de mineração de dados.

No entanto, procurou-se desenvolver a aplicação em uma arquitetura em camadas que, no futuro, permita facilmente adaptar o mecanismo para tratar outras fontes de informação, como, por exemplo, arquivos de um editor de texto, outras redes sócias etc. Para isso, o objetivo principal deste projeto consiste em criar um *framework* para realizar a indexação do conteúdo dos *tweets* de uma pessoa, de forma que essa base de dados possa servir de entrada para outro *framework*, que realiza a Recuperação da Informação e é responsável pela apresentação dos dados na forma de uma nuvem de *tags*, permitindo assim, a socialização do conhecimento contido nas bases de dados indexadas.

Ao tornar pública sua nuvem de *tags*, o usuário dá mais visibilidade à sua produção, mostrando ao público quais assuntos possui maior domínio e interesse. Além disso, pretendesse que a nuvem de *tags* funcione como um indexador que, ao ser escolhido um termo específico, retorna uma lista dos documentos relacionados com aquele termo.

Um fator motivador para esse projeto é o fato de, recentemente, o *Twitter* ter liberado um recurso para os usuários baixarem todos os seus *tweets* (FELIX, 2013). Até então, a empresa limitava a recuperação de um pequeno número de *tweets* (aproximadamente 1000 *tweets*) através da API de desenvolvimento fornecida para terceiros.

Agora, com esse recurso, o usuário pode fazer o *download* de todo o seu histórico de *tweets* uma única vez e montar a base de conhecimento inicial. A posterior atualização dessa base de conhecimento pode ser feita em períodos longos, por exemplo, a cada seis meses, através de novo *download* dos *tweets* do usuário.

Após o termino da etapa de implementação e da realização de testes da aplicação, pretende-se que a mesma seja disponibilizada na nuvem (*cloud computing*), de forma que possa tornar-se um produto de uso público.

PROCEDIMENTOS METODOLÓGICOS

O presente projeto de pesquisa, classificada como Pesquisa Aplicada, iniciou com a análise dos arquivos de *tweets*, disponibilizada através de *download* pelo *Twitter* por meio de arquivos textos no formato de documento JSON contendo o histórico do usuário. O objetivo desta análise foi identificar a estrutura desses arquivos e permitir a modelagem de um banco de dados para armazenar as informações necessárias para a aplicação. A análise da estrutura dos arquivos e dos conteúdos dos *tweets* também permitiu definir uma classificação dos dados, separando as diferentes informações, cada uma em seu respectivo campo.

A aplicação como um todo foi estruturada em camadas (Figura 1), de forma que a implementação de cada operação torne-se independente das demais, permitindo, assim, o reuso dos componentes de *software* produzidos em projetos futuros.

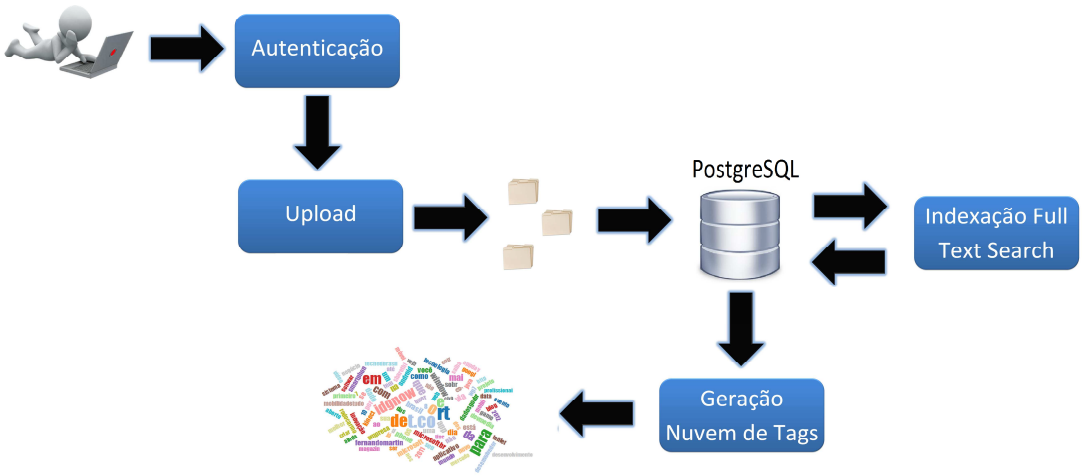

Figura 1 – Arquitetura da aplicação.

Visto isso, na sequência, foram desenvolvidas as camadas de *software* relativas ao processo de *upload* dos arquivos de *tweets*. Esta ferramenta faz o controle do usuário que está enviando os arquivos a partir de um *login* com uma conta do próprio *Twitter* (SOCIAL LOGIN, 2014). Com este sistema de *login*, pode-se controlar quem está enviando cada arquivo e coletar algumas informações sobre este usuário. Após fazer esse *login*, o usuário é redirecionado para uma página na qual ele deve anexar a sua base de dados já gerada anteriormente pelo *Twitter*.

A partir deste momento começa a ser feita a manipulação dos dados anexados. Para isso, é feita a classificação das informações, na qual se pega os dados presentes no arquivo e coloca-se nos respectivos campos criados no banco de dados. Os dados que estão sendo utilizados para o processo de geração da nuvem de *tags*, neste primeiro momento, são os referentes aos textos publicados pelo usuário em seus *tweets*. Eventualmente, outros dados de interesse podem surgir e, para isso, também está sendo armazenado no banco de dados o arquivo completo que é anexado pelo usuário.

As informações coletadas são utilizadas em um mecanismo de Recuperação da Informação, que extrai as palavras mais frequentes e os termos mais relevantes que estão contidos nos arquivos anexados, armazenando, ainda, o número de repetições

(frequência de citação) de cada um destes termos. Para extrair essas palavras foi utilizado o recurso de indexação textual do *PostgreSQL*. Esse mecanismo de indexação, denominado *Full Text Search*, realiza diversas funções como a remoção de *Stop Words* e utilização de lexemas, que é um conceito similar ao conceito de radical em gramática (POSTGRESQL, 2014). Para realizar a Recuperação da Informação foi criado um gatilho (*trigger*) que, ao realizar a inserção ou alteração de registros, é acionado e executa a indexação do texto, salvando-as em um campo especifico. Este campo contem os termos relevantes acompanhados pelo número de repetições e é utilizado pela camada de geração da nuvem de *tags*.

As informações extraídas pelo mecanismo de Recuperação de Informação são passadas para outro mecanismo, que é responsável por realizar a criação da nuvem de *tags*. Este mecanismo recebe uma lista com as principais palavras acompanhadas pelo número de vezes que cada uma delas aparece para, em seguida, realizar a distribuição das palavras, variando seu tamanho de acordo com o número de repetições, ou seja, quanto maior o número de repetições de uma palavra maior é seu tamanho na representação.

A camada final de apresentação permite que o usuário tenha uma representação visual, através da nuvem de *tags*, dos assuntos presentes em sua base de dados, com destaque aos termos que identifiquem os assuntos de maior frequência.

Percebe-se a importância desta pesquisa para a comunidade em geral, em especial, para aquelas pessoas que produzem algum tipo de conhecimento na forma de textos. Com isso, além de servir como mecanismo de indexação de suas produções, o sistema permite que o usuário dê maior visibilidade a sua produção, por meio da publicação de sua nuvem de *tags*, tornando-a disponível para qualquer usuário. O resultado do projeto compreende uma solução com aplicação prática na resolução de problemas associados com a Recuperação de Informação sobre a base de dados produzida pelo usuário. Mais especificamente, esse projeto limita-se à base de dados construída sobre os *tweets* de um usuário que, atualmente, conta apenas com a pesquisa sobre palavras chave.

RESULTADOS E DISCUSSÂO

A primeira camada de *software* que foi desenvolvida foi a camada relativa ao processo de identificação do usuário que anexa uma base de dados. Para isso, a aplicação redireciona o usuário para uma página do *Twitter* na qual ele deve identificar-se e autorizar que a aplicação tenha acesso ao seu cadastro no *Twitter*, realizando o *login* social. Após a autenticação, essa página do *Twitter* redireciona o usuário novamente para a aplicação, fornecendo agora um *token* que permite acessar algumas informações sobre este usuário. Essas informações são importantes para saber quem está enviando determinada base de dados e para que o usuário possa anexar mais do que uma base de dados, gerando assim, sua nuvem de *tags* baseada em todo o histórico que aquele usuário possui e para que, mais tarde, caso o usuário deseje, possa compartilhar sua nuvem de *tags* a partir da própria aplicação.

A segunda camada desenvolvida é aquela que o usuário usa para anexar a sua base de dados gerada anteriormente pelo *Twitter*. A base de dados deve estar em formato JSON (arquivos com extensão .js) e é a ultima etapa a ser feita pelo usuário.

Nesta etapa, o usuário escolhe cada arquivo do histórico e envia-o. Um usuário pode anexar várias bases de dados, uma vez que, para a geração da nuvem de *tags*, são pegos todos os arquivos anexados por este determinado usuário.

Na mesma tela, e ao lado dos campos para anexar os arquivos, tem-se um botão verde em que está escrito "*Gerar Nuvem*". Este botão é responsável por iniciar o processo que gera a nuvem de *tags* baseado nos arquivos enviados até o momento. Para gerar essa nuvem foram desenvolvidas mais duas camadas, uma responsável por realizar a Recuperação da Informação e outra responsável por criar a nuvem de *tags* com base nos dados passados pela camada anterior.

A camada que realiza a Recuperação da Informação ainda precisa ser melhorada em trabalhos futuros. Até o momento é utilizado o recurso de indexação textual do *PostgreSQL*, denominado *Full Text Search*. Este mecanismo possui algumas outras funções que ainda precisam ser analisadas e testadas. No caso da camada desenvolvida até o momento, foi criado um gatilho (*trigger*) que, ao realizar a inserção ou alteração de registros, é acionado e executa a Recuperação destas Informações, salvando-as em um campo especifico.

A camada responsável por realizar a criação da nuvem de *tags* recebe os valores obtidos após a Recuperação das Informações. Esses valores são os principais termos relevantes que aparecem na base de dados acompanhados do respectivo número de repetições. De posse destes valores, essa camada realiza a distribuição aleatória das palavras na nuvem, variando seu tamanho de acordo com o número de repetições - quanto maior for o número de repetições, ou seja, quanto mais vezes a palavra aparece na base de dados, maior é o tamanho de sua fonte na representação da nuvem. Além disso, as palavras são diferenciadas por cores aleatórias para facilitar a distinção dos termos. Na Figura 2 pode-se ver um exemplo de nuvem de *tags* gerada pela aplicação.

Figura 2 - Exemplo de nuvem de *tags* gerada.

Na Figura 2 pode-se ver, também, a necessidade de melhorar o mecanismo de Recuperação da Informação, o qual considera diversas palavras sem relevância, como “de”, “t.co”, “rt”, “8”, “dos”. A base de testes possui cinco arquivos com dados reais, que correspondem aos *tweets* do usuário de testes do período de jul./2011 a nov./2011, totalizando 183 *tweets*. Sendo que cada *tweet* pode ter até 147 caracteres.

CONSIDERAÇÕES FINAIS

Para trabalhos futuros espera-se melhorar a filtragem das *Stop Words*, retirando termos como partes de URLs, artigos, conjunções, preposições, pronomes e outras palavras sem relevância para a nuvem de *tags*. Também, como trabalho futuro, pretende-se realizar melhorias no *layout* do *site*, permitir a publicação da nuvem de *tags* gerada em redes sociais, adquirir um domínio próprio para o projeto e realizar mais algumas melhorias de acordo com analises feitas sobre o trabalho.

Um protótipo funcional da aplicação desenvolvida pode ser acessado através do *link* http://54.186.98.82/twittknowledge.

O presente trabalho foi realizado com apoio do CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil, através do Edital Nº 106/2013 PIBITI/PIBIC/PIBIC-Af/CNPq/IF CATARINENSE.

REFERÊNCIAS

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: Estadao, 18 jan. 2013.
MASHABLE. **Will You Be Twitter’s 500 Millionth User?** fev., 2012. Disponível em: <http://mashable.com/2012/02/22/twitters-500-million-user/>. Acessado em: 25 Jul. 2014.

POSTGRESQL. **Chapter 12. Full Text Search**. Disponível em: <http://www.postgresql.org/docs/8.3/static/textsearch.html>. Acesso em: 01 mar. 2014.

SOCIAL LOGIN. In: Wikipedia – The Free Encyclopedia. Disponível em: <http://en.wikipedia.org/wiki/Social_login>. Acesso em: 01 jul. 2014.

TEIXEIRA, F. A. G.; DUQUE, C. G. A Recuperação da Informação e a colaboração de usuários na Web – Novas oportunidades para a Comunicação. In: CONGRESSO INTERNACIONAL COMUNICACION 3.0, 3., Salamanca (ES), 2012. **Proceedings...** Salamanca: Universidad de Salamanca, 2012.