CONHECIMENTO EM REDES SOCIAIS: Um estudo relacionado à recuperação de informação sobre textos publicados no *Twitter*

Miguel Airton Frantz[1]; Angelo Augusto Frozza[2]; Reginaldo Rubens da Silva[3]

RESUMO

Este trabalho tem como objetivo apresentar um protótipo capaz de realizar recuperação de informações em textos de mensagens postadas por um usuário do *Twitter*. Com o intuito de permitir fácil adaptação da ferramenta para trabalhar com textos de outras fontes de dados, a mesma foi desenvolvida em camadas. Dessa forma, a aplicação extrai palavras chave encontradas nas mensagens postadas de cada usuário. O resultado é apresentado em forma de uma nuvem de *tags*, atribuindo maior tamanho de fonte aos termos relevantes que aparecem com mais frequência. A nuvem de *tags* resultante representa, assim, os assuntos mais abordados pelo usuário.

Palavras-chave: Recuperação da informação, *Twitter*, Redes sociais, Internet, *Web* 2.0.

INTRODUÇÃO

A *Web* 2.0, conhecida também como a *Web* Social, tem transformado a forma como as pessoas se relacionam na rede, bem como tem revolucionado o mundo do trabalho, desafiando os indivíduos e as organizações a aumentarem sua eficiência, principalmente na produção e disseminação de informação e de conhecimento. As tecnologias e aplicações da *Web* 2.0 são utilizadas tanto para o simples compartilhamento de informações e de conhecimento entre amigos, como também, para a gestão estratégica nas organizações. As ferramentas da *Web* 2.0 estão focadas na criatividade, na informação compartilhada e, acima de tudo, na colaboração (ANDRADE *et al*., 2011).

Atualmente, a *web* constitui o maior repositório de informações existentes no mundo, tornando-se um importante meio de comunicação e disseminação de conhecimento. Uma enorme quantidade de dados é produzida, armazenada e compartilhada diariamente. As informações estão dispostas de diversas maneiras na *web*, como em *blogs*, fóruns, redes sociais, *websites* ou outros meios, e compõem imensos bancos de dados, repletos de variadas informações. São tantas informações que o usuário acaba se perdendo e tendo dificuldade de encontrar o que realmente quer em meio a esta infinidade de conteúdo.

Tendo em vista este problema, surge a proposta de um mecanismo capaz de extrair as principais informações presentes em um conjunto de textos. Assim, é possibilitado ao usuário que indexe seus textos, para que a aplicação em questão, se encarregue de conceituar os assuntos tratados, resumindo de maneira clara e representativa os dados encontrados.

Com o intuito de facilitar a interpretação das informações obtidas pela aplicação, optou-se por apresentá-las no formato de uma nuvem de *tags*. A nuvem de *tags* consiste em um formato de representação visual que pode ser facilmente entendido, em que os termos relevantes encontrados na base de dados de um usuário, também conhecidas como *tags*, são dispostos formando uma nuvem. O tamanho de cada termo varia conforme o número de vezes que o mesmo aparece na base de dados – palavras mais frequentes recebem maior tamanho de fonte.

1 Estudante de Graduação em Sistemas de Informação, IFC-Camboriú. Bolsista do CNPq Brasil. *E-mail*: frantz.miguel@gmail.com.
2 Mestre em Ciências da Computação, UFSC; Professor do IFC-Camboriú. *E-mail*: frozza@ifc-camboriu.edu.br.
3 Mestre em Ciência e Tecnologia Ambiental, UNIVALI; Professor do IFC-Camboriú. *E-mail*: reginaldo@ifc-camboriu.edu.br.

Tendo em vista a diversidade de formatos de entrada de dados possíveis para esse mecanismo, optou-se por utilizar, neste primeiro momento, apenas textos de mensagens postadas por usuários do *Twitter*, que é uma das mais importantes redes sociais atualmente.

Segundo Benevenuto, Almeida e Silva (2012), as 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 ideal para extração de conhecimento e aplicação de técnicas de mineração de dados. 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).

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. Assim, o usuário pode fazer *download* de todo seu histórico de *tweets* uma única vez e montar a base de conhecimento inicial, atualizando sua base de dados conforme seu desejo, podendo ser em períodos curtos ou longos.

A função principal dos sistemas de Recuperação da Informação (RI) é organizar informações de forma ordenada e inteligente em bancos de dados. De forma geral, esses sistemas compõem outros grandes sistemas de informação, para aprimorar o processo de consulta aos dados contidos em seus repositórios (GOMES, 2011).

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

Na sequência, a seção Procedimentos Metodológicos apresenta a metodologia de desenvolvimento do projeto. Na seção Resultados e Discussão são discutidos os resultados obtidos até o momento. A seção Considerações Finais faz um encerramento do trabalho.

PROCEDIMENTOS METODOLÓGICOS

Dando início a este projeto, primeiramente realizou-se uma pesquisa bibliográfica sobre o tema. Para isso, realizaram-se buscas na Internet por artigos e trabalhos relacionados à recuperação da informação sobre textos do *Twitter*. Foram encontrados 22 artigos considerados relevantes para o projeto. Estes foram examinados mais detalhadamente e servem de fundamentação para as principais ideias e propostas apresentadas no decorrer deste artigo.

O objetivo desta pesquisa consiste em estudar o que está sendo feito e pesquisado no que diz respeito a Recuperação de Informação, a utilização do *Twitter* para essa tarefa, a utilização de *tags* e formas de representação destas *tags*, como é o caso da nuvem de *tags* que já é utilizada neste projeto.

Com a análise destes artigos foi possível identificar diversas técnicas de RI, apesar de grande parte dos trabalhos terem enfoques diferentes, e como elas podem ser encaixadas neste projeto.

RESULTADOS E DISCUSSÃO

Como resultado desta etapa do projeto, que consiste em uma revisão bibliográfica, tem-se um conjunto de informações e embasamentos sobre o tema. Assim, pode-se evidenciar e constatar a importância de ferramentas de Recuperação da Informação para a sociedade atual, tendo em vista a enorme quantidade de informações que são produzidas e que circulam diariamente pela Internet. Segundo Coelho (2014), com o advento da *Web* 2.0, novos produtores de conteúdo surgiram, o que levou à necessidade de dispositivos que representassem todo o conteúdo produzido para uma posterior recuperação através dos *sites* de busca. Assim, as *tags* e, mais recentemente, as *hashtags*, passaram a desempenhar esta função. Entretanto, devido ao ambiente heterogêneo que se constitui a Internet, torna-se impossível estabelecer regras que possam reger a produção e a recuperação informacional.

Classificação do Conhecimento

Segundo Rufino (2010), há muito tempo o homem separa e classifica tudo que existe, mesmo que isto ocorra de maneira inconsciente, como a classificação das roupas, sapatos, CDs, DVDs, fotos, cartas, livros, alimentos, entre outros. Seja por cor, tamanho, conteúdo ou importância, as pessoas estão sempre separando objetos. O mesmo acontece com o conhecimento. Diante do crescente aumento na produção de documentos e da participação cada vez mais frequente dos usuários no ambiente em que as informações são produzidas, surge uma nova necessidade. É preciso, além de classificar as informações rapidamente, permitir ao usuário uma interação com o sistema e com as informações disponíveis. Dessa necessidade, surgiu uma nova ideia de classificação, denominada *folksonomia*, baseada na filosofia colaborativa, emergente dos preceitos de interatividade da *Web* 2.0.

Segundo Catarino e Batista (2007),

> *Folksonomia* é a tradução do termo *folksonomy* que é um neologismo criado em 2004 por Thomas Vander Wal, a partir da junção de *folk* (povo, pessoas) com *taxonomy*. [...] *Folksonomia* é o resultado da etiquetagem dos recursos da *Web* num ambiente social (compartilhado e aberto a outros) pelos próprios usuários da informação visando a sua recuperação.

Sousa (2012) também diz que o ato de classificar acompanha o cotidiano dos seres humanos, muitas vezes de forma imperceptível. Ela cita a rede social *Twitter*, que utiliza um sistema de classificação e indexação opcional das mensagens, o qual consiste em atribuir uma *tag* (etiqueta, em português) 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 na rede social.

Costa e Rodrigues (2010) dizem que se está diante de uma questão que merece uma elaboração de hipóteses levando-se em consideração o crescimento exponencial da massa documental, as mudanças ocorridas nas necessidades e no uso da informação pelos usuários de Centros de Informação em geral e nos avanços tecnológicos que cada vez mais se expandem com maior velocidade, afetando diretamente a produção editorial de várias áreas do conhecimento.

Segundo Vance *et al*. (2009), o *Twitter* pode desempenhar um papel importante na gestão da informação. Ele permite identificar indivíduos, instituições e periódicos e, desta forma, atua como filtro de informação, pois, por meio das mensagens compartilhadas, é possível acessar diretamente a informação mais

relevantes para uma determinada área de interesse. Agências internacionais, como a OMS <@whonews> ou o Centro Controle de Doenças dos Estados Unidos <@CDCgov>, utilizam o *Twitter* para divulgar recomendações, surtos de doenças e novas estratégias de prevenção.

Recuperação da Informação em redes sociais

De acordo com Costa e Rodrigues (2010), os Sistemas de Recuperação da Informação (SRIs) objetivam organizar, registrar e fornecer acesso às informações contidas em documentos de determinado acervo. Exercem papel fundamental na organização e representação da informação para a recuperação.

Gomes (2011) diz que a função principal dos sistemas de RI é organizar informações de forma ordenada e inteligente em bancos de dados. Via de regra, esses sistemas compõem outros grandes sistemas de informação, para aprimorar o processo de consulta aos dados contidos nesses repositórios.

Segundo Manning e Schütze (2009), as operações básicas dos sistemas de Recuperação da Informação (RI) consistem de quatro etapas, sendo elas: a análise léxica; a remoção de termos comuns ou *Stop Words*; a normalização, que é a classificação de equivalência dos termos com criação de grupos nominais; e a fase de *stemming* e *lemmatization*.

Na etapa da análise léxica é realizada a remoção dos caracteres indesejados presentes no texto, os quais são definidos de acordo com o contexto de cada aplicação. Na etapa da remoção de *Stop Words*, são removidos os termos cuja ocorrência é bastante comum em qualquer texto. São os pronomes, verbos, advérbios, artigos e outros termos dotados de pouco valor semântico, sendo que, da mesma forma que ocorre na análise léxica, os critérios para a remoção dependem da necessidade de cada aplicação. A identificação de grupos nominais consiste na distinção, em meio aos diversos termos, de grupos de palavras como "Sistemas de Informação", "Banco de Dados" ou mesmo, "Recuperação de Informação". Na fase de *stemming*, a ideia é reduzir os termos aos seus respectivos radicais. Para isso, é realizado o agrupamento de diferentes formas de uma palavra, como é o caso de organizar, organize, organizando (GOMES, 2011).

Existem diversas abordagens na literatura para realizar esta tarefa de forma semelhante. Em geral, o objetivo é realizar uma redução do léxico, que 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 seu sentido. Esses termos podem ser entendidos como as palavras chave do texto em questão e são elas que são utilizadas na aplicação desenvolvida neste projeto.

TweetKnowledge

O *TweetKnowledge* consiste em uma aplicação prática, desenvolvida na primeira etapa do presente projeto, para a resolução de problemas associados com Recuperação da Informação sobre uma base de dados produzida por um usuário ao compartilhar notícias e informações por meio de sua conta do *Twitter*. Até o momento, o projeto limita-se a essa base de dados, mas como já mencionado anteriormente, buscou-se desenvolver a aplicação em camadas para facilitar sua reutilização com outros tipos de informações. O resultado final da aplicação desenvolvida consiste em uma nuvem de *tags*, conforme pode ser visto na Figura 1.

FIGURA 1 – Nuvem de *tags* gerada pelo *TweetKnowledge*.

Atualmente, estão sendo estudadas outras alternativas para a realização da indexação dos textos dos *tweets*, a fim de obter melhor qualidade na limpeza dos dados e, por consequência, maior exatidão na seleção dos termos relevantes dos textos. Além disso, também está sendo proposto utilizar *Linked Data* para adicionar conhecimento aos termos da lista. Assim, cada termo mostrado na nuvem de *tags* apontará para os textos (*tweets*) de sua origem e também para fontes externas, com o intuito de adicionar mais informações e o conhecimento propriamente dito sobre aquele termo em específico.

CONSIDERAÇÕES FINAIS

Tendo em vista a importância da Recuperação da Informação em redes sociais e também a relevância da classificação do conhecimento utilizando *tags* e palavras chave que representem uma mensagem ou texto, o presente projeto disponibiliza um mecanismo para extração de conhecimento em cima dos *tweets* de um usuário e permite a representação das informações obtidas de maneira que o usuário possa entender com facilidade.

Para tanto, é oferecida uma ferramenta que realiza a busca de palavras chave presentes em todo o histórico de publicações no *Twitter* que um usuário desejar. 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, esses valores são passados para a camada responsável pela geração da nuvem de *tags*, que realiza a distribuição aleatória das palavras na nuvem, variando seu tamanho de acordo com o número de ocorrências de determinado termo na base de dados. Quanto maior o número de ocorrências, maior é seu tamanho na representação da nuvem de *tags*, dando assim, maior destaque aos assuntos

(termos) que são mais frequentes. Além disso, as palavras são diferenciadas por cores aleatórias para facilitar a distinção dos termos.

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 a proposta de utilizar *Linked Data* para adicionar conhecimento na lista de termos, sendo que cada termo da nuvem de *tags* direcionará para os textos (*tweets*) de sua origem e também para fontes externas.

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

REFERÊNCIAS

ANDRADE, I. A. *et al*. Inteligência coletiva e ferramentas *Web* 2.0: a busca da gestão da informação e do conhecimento em organizações. **PG&C - Perspectivas em Gestão & Conhecimento**. João Pessoa, v. 1, Número Especial, p. 27-43. 2011.

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.

CATARINO, M. E.; BAPTISTA, A. A. Folksonomia: um novo conceito para a organização dos recursos digitais na Web. **DataGramaZero: Revista de Ciência da Informação**, v. 8, n. 3, jun. 2007.

COELHO, V. L. **Hashtags**: Rompimentos com dizeres sedimentados. XI EVIDOSOL e VIII CILTEC-Online. In: <http://evidosol.textolivre.org>. 2014.

COSTA, F. P. da; RODRIGUES, L. S. A democratização e a indexação da informação: como wikis e folksonomias podem afetar o âmbito informacional. In: CONGRESSO NACIONAL UNIVERSIDADE, EAD E SOFTWARE LIVRE – UEADSL. **Anais…** 2010.

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.

GOMES, L. F. A. **Nuvens de tags no Twitter**: estudo e implementação. In: <http://www.cin.ufpe.br/~tg/2011-1/lfag.pdf> Recife, 2011.

MANNING, D.; RAGHAVAN, P.; SCHÜTZE, H. **An Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2009.

RUFINO, A. Folksonomia: novos desafios do Profissional da informação frente às novas possibilidades de organização de conteúdos. **Múltiplos Olhares em Ciência da Informação**, v. 1, n. 1, out. 2010.

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

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.

VANCE, K.; HOWE, W.; DELLAVALLE, R. P. Social internet sites as a source of public health information. **Dermatologic Clinics**, v. 27, n. 2, p.133-136, 2009.