13 A 15 DE AGOSTO DE 2014
III WPCI
Workshop de Pesquisa em Ciência da Informação
"ABORDAGENS CONTEMPORÂNEAS NA CIÊNCIA DA INFORMAÇÃO"

Eixo Temático: Compartilhamento da Informação e do Conhecimento

COLETA DE DADOS EM PLATAFORMAS DE REDES SOCIAIS: ESTUDO DE APLICATIVOS

Katia Maria Poloni[1]
Maria Inês Tomaél[2]

RESUMO
O compartilhamento da informação necessita ser estudado no âmbito das mídias sociais, para isso o objetivo deste trabalho foi identificar aplicativos de acesso livre que fazem extração de dados em plataformas como Facebook, Flickr, YouTube e Twitter. Buscou-se aplicativos, por meio da literatura e da *Web*, que disponibilizam recursos para gerar grafos de rede e para subsidiar as análises de redes sociais. O aplicativo Netvizz foi selecionado para análises no Facebook, o Remid para o twitter e plugin NodeXL, que possui seu próprio gerador de grafos, foi selecionado para o Flickr, YouTube e Twitter. Para as demais mídias sociais, foi adotado o programa Gephi. A funcionalidade dos aplicativos depende da mídia social a ser estudada, cada mídia possui política de privacidade que pode limitar a extração de dados.

Palavras-chave: Mineração de dados. Visualização de grafos. Mídias sociais. Análise de Redes Sociais. Compartilhamento da informação.

1 INTRODUÇÃO E OBJETIVOS

Uma rede social representa uma estrutura social composta por pessoas ou organizações, conectadas por um ou vários tipos de relações, que partilham valores e propósitos comuns (FERREIRA, 2011). As plataformas de redes sociais na Internet possibilitam o compartilhamento de informações por meio das ligações de atores que essas mídias sociais possibilitam.

As mídias sociais permitem que pessoas com objetivos, idades e crenças diferentes se comuniquem e troquem informações sobre diversos assuntos – o público tornou-se livre para compartilhar suas opiniões e sentimentos; ou simplesmente informação. O compartilhamento da informação pode gerar conhecimento e é importante que seja estudado. De acordo com Rezende (2005), tal conhecimento pode ajudar na análise de padrões que podem ajudar na previsão de fatos futuros, condição que corrobora com este estudo, que apresenta aplicativos para a extração de dados de mídias sociais, resultante do compartilhamento de informação entre atores de uma rede.

[1] Graduanda do Curso de Ciência da Computação – UEL. Email: katiampoloni@gmail.com
[2] Professora do Departamento de Ciência da Informação – UEL. Email: maritomael@gmail.com

O compartilhamento de informação constitui-se na troca de informações entre os parceiros, que produzem o aumento da visibilidade da cadeia que abastece os processos nos quais estão inseridos. Wang et al. (2008) acreditam que o compartilhamento de informações com parceiros pode não só aumentar os benefícios operacionais, mas também os benefícios estratégicos de uma organização. Além disso, o compartilhamento de informações pode ser visto como um sinal de interesse de uma organização em construir um relacionamento de longo prazo e, desse modo, fortalecer a parceria.

Para possibilitar análises diversificadas do compartilhamento da informação em mídias sociais, faz-se necessário a coleta de dados veiculados nessas mídias. Esta pesquisa tem o objetivo de identificar aplicativos de acesso livre que fazem extração de dados em plataformas como Facebook, Flickr, YouTube e Twitter, e, mais especificamente: a) Levantar aplicativos para a coleta de dados em plataformas de redes sociais; b) Selecionar aplicativos que identifiquem as ligações entre os membros de uma rede; c) Testar estes aplicativos para verificar sua usabilidade; d) Identificar aplicativos que coletam dados de ligações, entre os membros de uma rede.

2 PRESSUPOSTOS TEÓRICOS

Para maior entendimento do ambiente virtual apresentamos conceitos de dois termos abordados por Lévy (1998): cibercultura e ciberespaço. Segundo Lévy (1998) a definição destes dois termos é dada como:

> "[...] O ciberespaço (que também chamarei de "rede") é o novo meio de comunicação que surge da interconexão mundial dos computadores. O termo significa não apenas a infraestrutura material da comunicação digital, mas também o universo oceânico de informações que ela abriga, assim como os seres humanos que navegam e alimentam esse universo. Quanto ao neologismo "cibercultura", especifica aqui o conjunto de técnicas (materiais, e intelectuais), de práticas, de atitudes, de modos de pensamento e de valores que se desenvolvem juntamente com o crescimento do ciberespaço." (LÉVY, 1998, p.13).

Apesar do caráter volátil das informações veiculadas na Web, o ciberespaço tem se tornado um ambiente informacional de grande importância para o grande público, especialmente para profissionais e pesquisadores. O uso de diversas mídias e de informações de todos os tipos disponíveis pelos recursos Web criou o que Lévy

(1998) denomina de cibercultura. Na visão do autor, cibercultura constitui-se em um novo tipo de relacionamento social que os indivíduos mantêm com as tecnologias digitais disponíveis na Internet. A cibercultura envolve conceitos voltados a sociedade, a cultura e as tecnologias digitais. Colaborações sócio-culturais que promovem "[...] agenciamentos sociais das comunidades no espaço eletrônico virtual, ou, ciberespaço. [...] a cibercultura é a cultura contemporânea fortemente marcada pelas tecnologias digitais" (LÉVY, 1998, p.13).

O ciberespaço está em constante desenvolvimento, espaço em que, na atualidade, os usuários podem usufruir e compartilhar informação ("Web 2.0"). De acordo com Kaplan e Haenlein (2010), as mídias sociais foram construídas sobre os alicerces tecnológicos e ideológicos da Web 2.0, e permitem a criação e troca de informações geradas pelo usuário. A expressão foi utilizada pela primeira vez em 2004 para descrever um novo modo para utilizar a *World Wide Web* refere-se a uma plataforma em que o conteúdo e os aplicativos são continuamente modificados por todos os usuários de uma forma participativa e colaborativa.

As redes sociais, de acordo com Marteleto (2001, p.72) são "[...] um conjunto de participantes autônomos, unindo ideias e recursos em torno de valores e interesses compartilhados".

De acordo com Santos (2010), a web é atualmente o maior repositório de informações existentes no mundo. Na web pessoas interagem todos os dias e compartilham um grande volume de dados que se perdem em meio a tanta informação. No mesmo ângulo Rezende (2005, p.397) afirma "devido à incapacidade do ser humano de interpretar tamanha quantidade de dados, muita informação e conhecimento, possivelmente úteis, podem estar sendo desperdiçados, ficando ocultos dentro das Bases de Dados espalhadas pelo mundo".

Para recuperar tamanha quantidade de dados, e desse modo possibilitar a análise do comportamento de indivíduos e sua evolução dentro de uma rede social, é preciso que técnicas computacionais de mineração de dados sejam utilizadas (CERVI, 2008). A análise do comportamento de indivíduos, bem como de sua trajetória em uma rede social depende de técnicas computacionais para a mineração de dados. Circunstância que possibilitou a união entre as áreas das Ciências Sociais e da Computação.

No Brasil, a Análise de Redes Sociais (ARS) aplicadas à mídias sociais como o Twitter e o Facebook ainda é incipiente, tendo em vista que estas mídias sociais passaram a ser utilizadas apenas mais recentemente. Adicionalmente, muitas métricas, aplicadas a dados científicos – disponíveis na literatura – não são apropriadas para a extração e mineração de dados em mídias sociais. Isto se deve ao fato do conteúdo conter muitas gírias, erros ortográficos, diferentes sinais (como, por exemplo, os que identificam risos) e abreviações para representar ações e reações dos usuários.

Desse modo, este estudo pretende selecionar e apresentar aplicativos que possam ser utilizados para a extração de dados em mídias sociais e disponibilizar subsídios para o entendimento dos recursos e limitações que as próprias mídias oferecem.

3 PROCEDIMENTOS METODOLÓGICOS

Os procedimentos metodológicos incluíram: a) Levantamento e fichamento da literatura pertinente; b) Levantamento de aplicativos na Internet, na literatura e com especialistas no âmbito da UEL; c) Análise e testes dos aplicativos nos recursos de plataforma de redes sociais; e) Sistematização, análise e descrição das informações/dados coletados.

Os aplicativos foram selecionados ao longo primeiro semestre de 2014, a partir dos disponíveis na Web, na literatura científica e com especialistas. Foram realizados testes de instalação, configuração e utilização com o propósito de avaliar a consistência dos mesmos na mineração de dados. Os aplicativos foram testados no âmbito de um projeto de iniciação científica, por uma das autoras desse estudo e foram escolhidos com base em sua facilidade de uso e abrangência de recursos.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Os programas selecionados com base na avaliação de suas funcionalidades, como facilidade de utilização e na forma como eles lidam com as limitações, que são provocadas pelas mídias as quais ele atende. Após testes e seguindo a recomendação de Rosa, Silva e Silva (2012, p.5) – "... existe um aplicativo chamado Netvizz, que gera um grafo de toda a sua rede social incluindo os "pesos" das relações para geração de grafos", o aplicativo Netvizz foi selecionado por aproveitar

bem os recursos que o Facebook proporciona através de sua API (Graph API) - *Application Programming Interface*. Mantendo o mesmo procedimento proposto por Recuero e Zago (2012), o *plugin* para o Excel (2007/2010), NodeXL demonstrou ter mais recursos embutidos além de ser extrator de dados do Twitter, Flickr e YouTube. Esta ferramenta também gera seus próprios grafos da rede. Após conseguir cumprir o propósito com os dois programas (Netvizz e NodeXL), analisamos o Remid[3], que ganhou espaço com sua API Twitter4J, utilizada por Magalhaes et al. (2012, p.128) que destacam a "... biblioteca Twitter4J, que possibilita a integração da linguagem com o Twitter". Por ser um programa produzido no Departamento de Computação da UEL, ele permite alterações de código e melhor manipulação de seu banco de dados, fazendo com que aplicações mais específicas, como uma busca incessante de dois meses, possa ser realizada.

Na realização deste trabalho diferentes aplicativos gratuitos[4] foram avaliados e os programas foram obtidos após uma longa pesquisa de suas funcionalidades e limitações. As mídias sociais utilizam-se de A*pplication Programming Interface* (APIs) específicas, que oferecem um conjunto de rotinas e padrões estabelecidos por um software para a utilização das suas funcionalidades por aplicativos que não pretendem envolver-se em detalhes da implementação do software, mas apenas usar seus serviços. Para ter acesso ao banco de dados das mídias sociais, é necessário que seja criado um aplicativo no site da mídia, normalmente localizado na parte de desenvolvedor onde o site libera um *token* de acesso. A API do Twitter utilizada é a twitter4j, que é uma biblioteca Java não oficial para a API do Twitter. Que dá acesso a 1% de seu banco de dados sem restrições de amizade. A API de Facebook (Graph API) dá acesso somente a informações de amigos ou pessoas que permitirem que o aplicativo faça uso de suas informações. Devido às limitações das APIs os programas possuem variações de acordo com as peculiaridades para a extração de dados de cada mídia analisada. Os programas foram selecionados por serem completos ao utilizar bem as funcionalidades de suas APIs, de fácil instalação e utilização.

[3] Programa produzido sob supervisão do Prof. Dr. Sylvio Barbon Júnior.
[4] Dentre os quais: TouchGraph, PDI (Pentaho Data Integration) da suíte do Pentaho e YourTwapperKepper.

O aplicativo Netvizz extrai as conexões de uma determinada rede pessoal. Este aplicativo considera um grupo de amigos como o domínio total e relaciona as amizades em comum, relaciona os amigos com base nas páginas que curtem, agrupa os dados para amizades e interações entre grupos, trata as páginas como um usuário e vê as relações de curtir entre outras páginas e também cria redes para a interação dos usuários em torno das páginas. Para exemplificar, apresentamos a figura 1, que foi gerada pelo Gephi, com base em dados extraídos do Netvizz. A figura representa as ligações de amizade de um perfil pessoal do Facebook, em que ele considerou todos os amigos do perfil como um domínio e relacionou as amizades em comum.

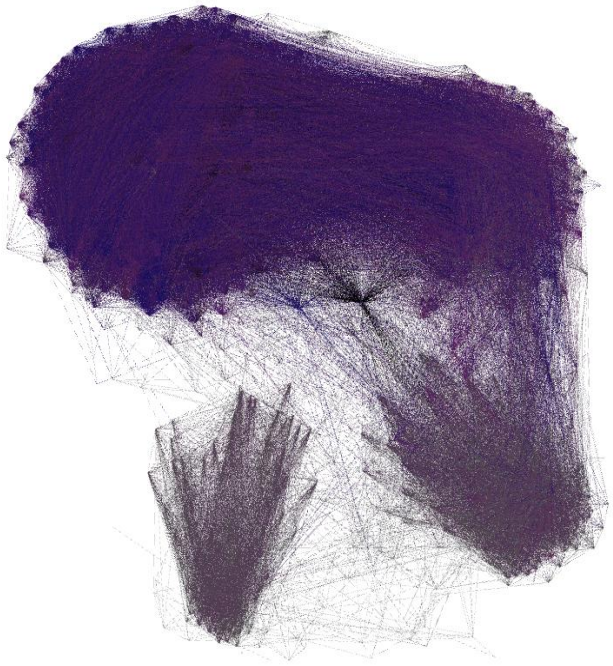

Figura 1: Grafo direcional do Facebook

O NodeXL é um *plugin* de acesso livre que funciona junto ao Excel versões 2007 ou 2010. Ele faz extração e análise de redes sociais como o Twitter, Flickr e YouTube. A extração pode ser por uma palavra-chave ou por um usuário no caso do twitter, usuário ou vídeo no caso do YouTube e por *tags* ou usuários no caso do Flickr. Ele produz também seu próprio criador de grafos com várias opções de edição. A Figura 2 exemplifica um grafo direcional formado por 100 tweets com o filtro "copa". Ele relaciona a citação (@nome usuário) de usuário, e caso não haja, direciona ao próprio remetente.

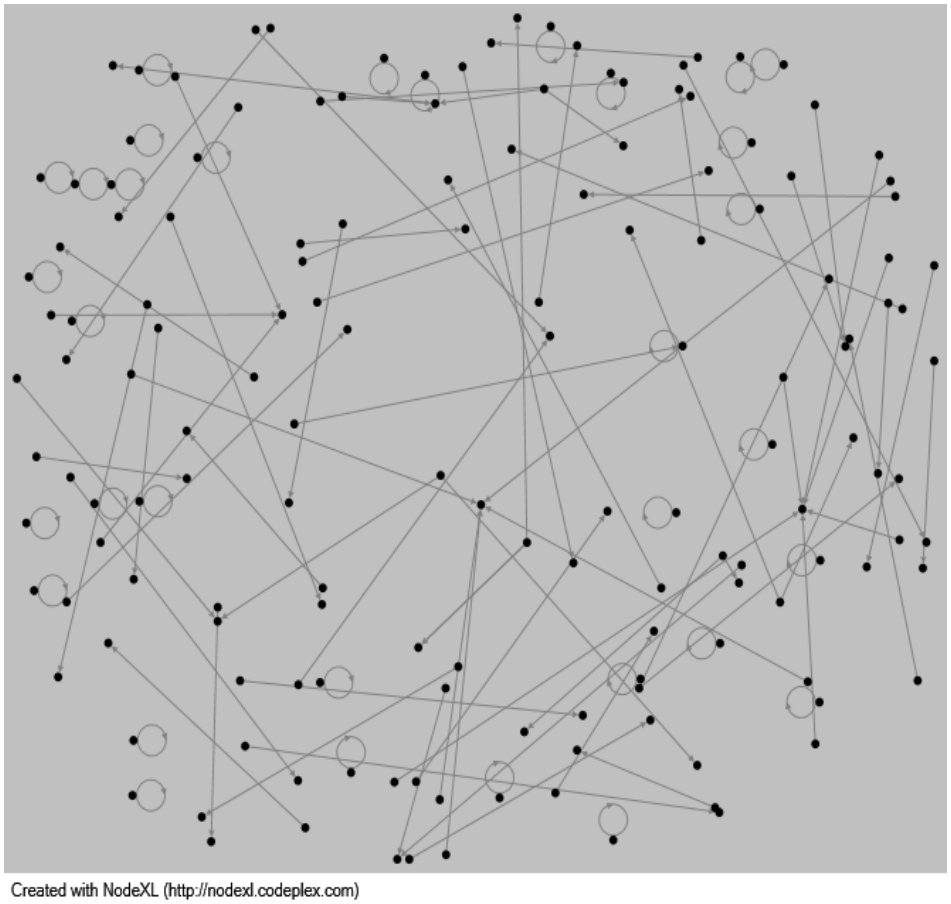

Figura 2: Grafo direcional do Twitter

O programa desenvolvido pelo grupo Remid da Ciência da Computação da UEL, faz extração de dados do twitter pela API twitter4j e coleta 1% dos dados em tempo real por meio de palavras-chave. Ele retorna o tweet completo, com informações de local, '#', citações (@), id do tweet, data e hora. A figura 3 ilustra um banco de dados dos tweets retornados.

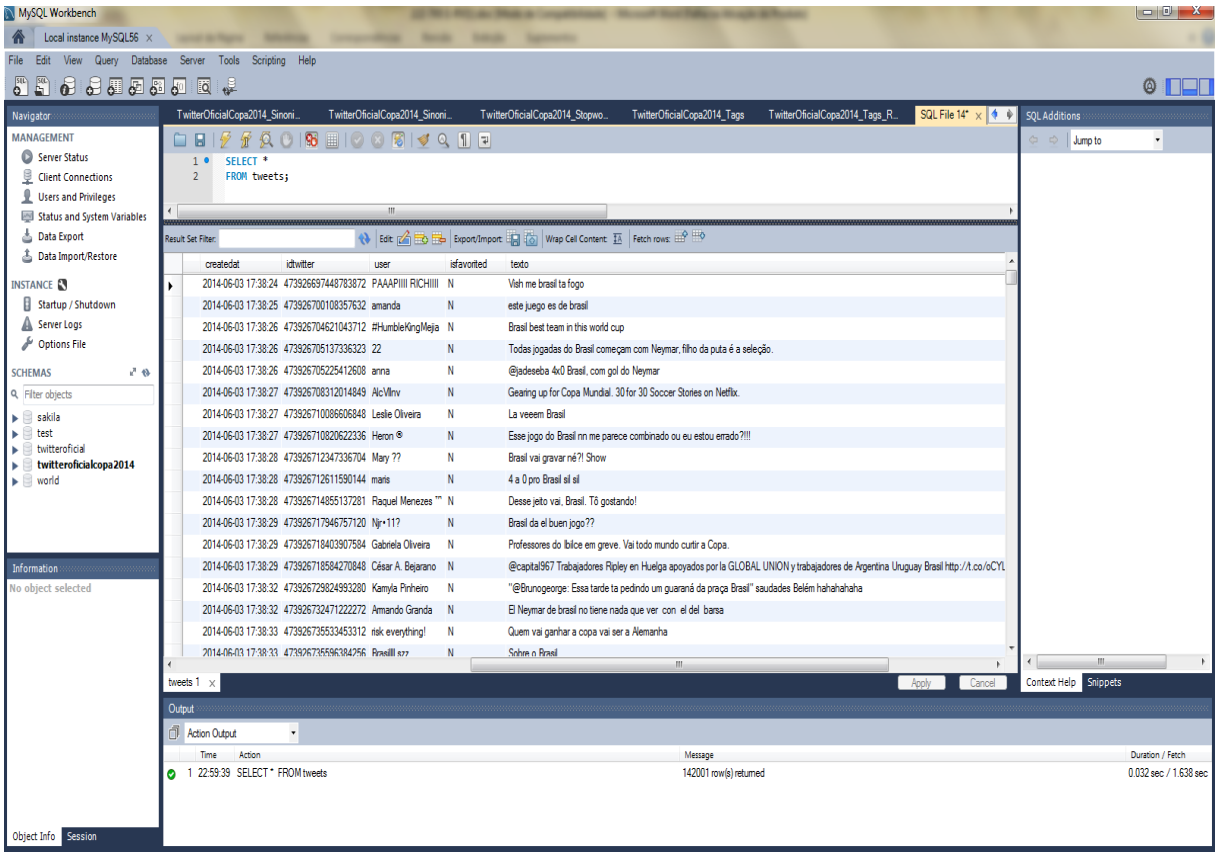

Figura 3: Banco de Dados de saída de uma extração do Twitter

O Remid possibilita a extração por até 10 palavras-chave do texto e conforme demonstra a figura 3, retorna os dados: data de postagem, identificador do tweet (numérico), usuário cadastrado no Twiiter e o texto postado no momento em que ocorre a extração dos dados.

A Figura 4 mostra um gráfico produzido no Gephi com 100 relações de *hashtags* utilizadas com o filtro "copa do mundo fifa" e em versão para a língua inglesa "*fifa world cup*", a extração foi feita no dia 25 de junho de 2014.

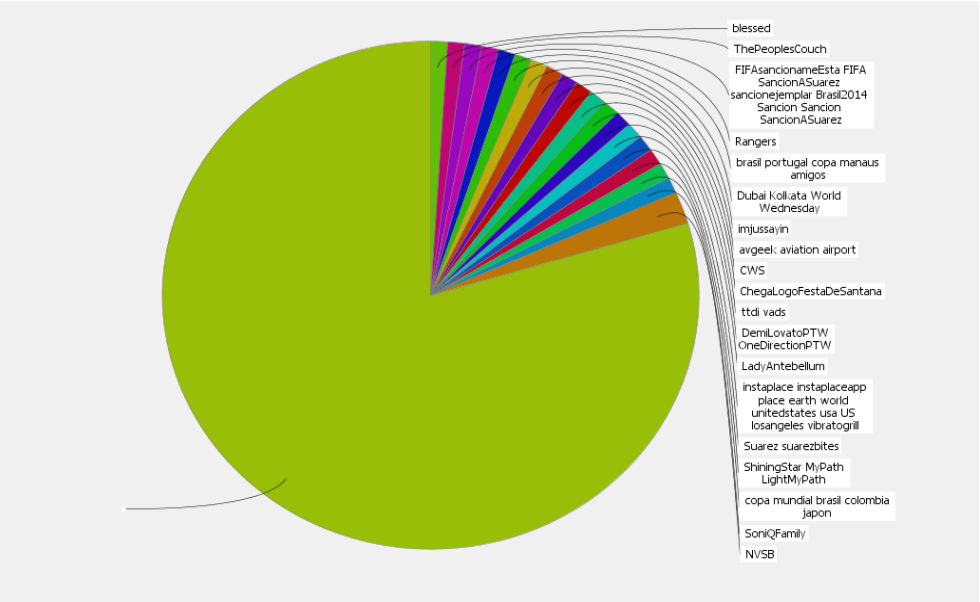

Figura 4: Hashtags(#) citadas no Twitter

Ao inserir os dados no Gephi, o gráfico acima foi elaborado com base nas associações por hashtags obtidas de 100 tweets. Nota-se que mais de 75% não fez uso do caractere (#), e os demais mantiveram-se bem distribuídos, sem concorrência.

5 CONSIDERAÇÕES FINAIS

As mídias sociais são recursos inovadores da contemporaneidade que precisam ser estudadas para o desvelamento das relações de atores e do compartilhamento da informação. A análise dos aplicativos demonstrou a facilidade de buscar dados e manipulá-los com os aplicativos selecionados. Embora algumas plataformas possuam certas limitações, não observamos nenhuma dificuldade de utilização para a realização dos testes. Observamos que os aplicativos estudados podem ser aplicados em diferentes pesquisas, sem a necessidade do apoio de um profissional com conhecimento técnico em Ciência da Computação.

Para extração de dados no Twitter, os programas com mais sucesso foram o NodeXL e o Remid da UEL. Ambos fazem a busca por palavras-chave e permitem criar um banco de dados para o pesquisador. O NodeXL foi selecionado para o Flickr e o YouTube, por gerar visualizações de gráficos de rede e por possibilitar a importação de dados para o Ucinet, Pajek e outros *softwares* para ARS.

Já para a extração de dados no Facebook, o programa escolhido foi o Netvizz, que considera um grupo de amigos como o domínio total e relaciona as amizades em comum, relaciona amigos com base nas páginas que curtem, agrupa os dados para amizades e interações entre grupos, trata as páginas como um usuário e vê as relações de curtir entre outras páginas e também cria redes para a interação dos usuários em torno das páginas.

A funcionalidade dos aplicativos depende da mídia social da qual os dados serão analisados, cada mídia possui uma política de privacidade que limita a extração de dados. É importante, entender as limitações de cada rede antes de fazer extrações, pois pode se tornar frustrante depender de dados que não serão adquiridos.

REFERÊNCIAS

CERVI, C. R. *Um Estudo sobre Mineração de Dados em Redes Sociais*. Porto Alegre: UFRGS, 2008. Dissertação (Mestrado em Computação) – Universidade Federal do Rio Grande do Sul.

MARTELETO, R.M. Análise de redes sociais: aplicação nos estudos de transferência da informação. *Ciência da Informação*, v.30, n.1, p.71-81, 2001.

KAPLAN, Andreas M.; HAENLEIN, Michael. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, n.53, p.59-68, 2010.

LÉVY, Pierre. *O que é virtual?* São Paulo: Editora 34, 1996.

MAGALHÃES, C. V. et al. Proposta de um Data Mart para avaliação de empresas usuárias do Twitter através das mensagens postadas pelos clientes. *Revista Brasileira de Administração Científica*. Aquidabã, v.3, n.2, p.123-135, 2012.

MANGOLD, W. Glynn; FAULDS, David J. Social media: The new hybrid element of the promotion mix. *Business Horizons*, v. 52, p.357-365, 2009.

RECUERO, Raquel; ZAGO, Gabriela. A Economia do Retweet: Redes, Difusão de Informações e Capital Social no Twitter. *Contracampo*, Niterói, v. 24, n. 1, p.19-43, jul. 2012.

REZENDE, S. O. Mineração de Dados. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 5. *Anais...* São Leopoldo: UNISINOS, 2005.

ROSA, A. C.; SILVA B. D.; SILVA P. L. Análise das redes sociais aplicada à engenharia social. In: SIMPÓSIO INTERNACIONAL DE GESTÃO DE PROJETOS, 1. *Anais...* São Paulo: UNINOVE, 2012.

SANTOS, Augusto Dias Pereira dos. *Descobrindo eventos locais utilizando análise de séries temporais nos dados do twitter*. Porto Alegre: UFRGS, 2013. Dissertação (Mestrado em Computação) - Universidade Federal do Rio Grande do Sul.

WANG, Chia-Chen et al. *Why Focal Firms Share Information*? A Study of the Effects of Power and Information Technology Competency. In: PACIFIC ASIA CONFERENCE ON INFORMATION SYSTEMS (PACIS). *PACIS 2008 Proceedings…* [S.L.]: AISeL, 2008. Disponível em: http://aisel.aisnet.org/pacis2008/68 Acesso em: 18 jul. 2014.

XAVIER T. C. *Estudo e desenvolvimento de jogos para internet utilizando Unity 3D*. Passo Fundo: UFRGS, 2011. Monografia (Especialização em Tecnólogo em Sistemas para Internet) – Universidade Federal do Rio Grande do Sul.

FERREIRA, Gonçalo Costa. Redes Sociais de Informação: uma história e um estudo de caso. *Perspect. ciênc. inf.* [online]. 2011, vol.16, n.3, pp. 208-231.