MINERAÇÃO DE DADOS DO TWITTER ATRAVÉS DO R

Mayara Kaynne Fragoso Cabral[1], João Carlos Bruno[2], Vanessa Aires Corado[3]

[1]Especialista em Banco de Dados e Professora do Instituto Federal Campus Porto Nacional – IFTO. e-mail: mayarakf@ifto.edu.br
[2]Graduando em Direito – Faculdade Serra do Carmo. e-mail: bruno.claroempresas@gmail.com
[3]Graduanda do curso de Licenciatura em Computação do Instituto Federal Campus Porto Nacional - IFTO. e-mail: vanessaa.corado@gmail.com

Resumo: Bilhões de pessoas criam trilhões de conexões através da mídia social a cada dia, cada clique constrói relacionamentos que, em conjunto, formam uma vasta rede social. Através da linguagem estatística R, este artigo apresenta o uso da mineração de dados para análise de dados sociais gerados na rede Twitter a partir do debate acerca da Redução da Maioridade Penal. A partir da hashtag #MaioridadePenal gerou-se inferências sobre os dados e mapeamentos gráficos que demonstraram posicionamentos contrários e favoráveis e o potencial tanto do R como da mineração de textos oriundos de redes sociais.

Palavras–chave: Twitter, Redução da Maioridade Penal, Mineração de Dados

1. INTRODUÇÃO

As redes sociais na Internet representam um novo e complexo universo de fenômenos comunicativos, sociais e discursivos [Recueiro 2014]. Não porque sua análise seja algo novo, mas porque a linguagem do ciberespaço tece hoje um caminho de amplos sentidos e múltiplas possibilidades [França 2009].

As consequências de um pensar mais global representado pela Internet e sua incidência sobre grande parte da vida da população, já eram refletidas no início do século por [Castells 2003] e [Lévy 2001]. Porém até os últimos 40 anos mais ou menos, a opinião sobre mineração quase não existia [Danneman e Heimann 2014]. Hoje a análise do comportamento de indivíduos, bem como de sua trajetória em uma rede social depende de técnicas computacionais para a mineração de dados [Poloni e Tomaél 2014].

As técnicas tradicionais de modelagem em ciências sociais tendem a necessitar de conjuntos de dados em que observações são independentes um do outro. [Danneman e Heimann 2014]. Em redes sociais os dados não são extraídos em uma amostra aleatória, e sim a partir de observações a respeito de um determinado assunto, ou seja, a partir de folksonomia.

Este termo segundo [Wal 2005] é o resultado da atribuição livre e pessoal de etiquetas (*tagging*) a informações ou objetos, visando à sua recuperação. Segundo Ji et al. (2007) etiquetagem colaborativa é um instrumento importante para classificar o conteúdo dinâmico para a pesquisa e compartilhamento.

Em plataformas como o *Twitter*, os usuários tendem a expressar livremente através das *hashtags*, o que cria um meio ideal de se capturar as opiniões comuns sobre diversos tópicos. Esta informação, no entanto, acaba por se tornando muito extensa inviabilizando que a análise na íntegra de todas as opiniões expressadas. Neste cenário o uso de ferramentas automáticas capazes de extrair o sentimento geral contido nos dados se torna muito atrativo.

Com o objetivo de construir uma melhor visão das possibilidades sobre a mineração de dados sociais a partir de rótulos utilizados em debates nas redes sociais, em especial no *Twitter*, realizou-se uma investigação em torno da *hashtag* #MaioridadePenal, que foi usada para um debate polêmico acerca da redução da maioridade penal no dia 08 de abril de 2015, data da instalação de uma comissão especial na Câmara dos Deputados para analisar a Proposta de Emenda Constitucional (PEC) para redução da maioridade penal de 18 para 16 anos.

ISSN 2179-5649
VI JICE©2015

O artigo se propõe a identificar o impacto da propagação de uma ideia por meio de uma *hashtag* no *Twitter* através das funções de mineração de textos da linguagem de programação estatística R (R Development Core Team, 2013). Neste estudo de caso espera-se minerar opiniões e posicionamentos contrários e favoráveis, permitindo uma analise do contexto social em torno do debate sobre a maioridade penal.

Este artigo se organiza em cinco seções. Na seção dois é realizado o levantamento de materiais bibliográficos e o embasamento conceitual, na seção três é apresentado o procedimento de mineração de dados utilizado para a construção da pesquisa e análise dos dados; na seção quatro são apresentados os resultados e discussões e na seção cinco as considerações finais da pesquisa realizada.

2. CONCEITOS

2.1 Folksomias e Redes Sociais

Conectividade e mídias sociais alteraram a forma como organizamos nossa comunicações. Bilhões de pessoas criam trilhões de conexões através da mídia social a cada dia, cada clique constrói relacionamentos que, em conjunto, formam uma vasta rede social [Hansen, Shneiderman e Smith, 2011].

O termo 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]. São criadas a partir da coleção de conexões entre um grupo de pessoas e coisas [Hansen et.al 2011 p. 4] e se tornam assim um conjunto de participantes autônomos, unindo ideias e recursos em torno de valores e interesses compartilhados [Marteleto 2001].

Segundo Golder e Huberman (2006) embora a organização eletrônica de conteúdo desta forma não é nova, uma forma de colaboração deste processo, que foi dado o nome de "etiquetar" por seus proponentes, é ganhando popularidade na web. A marcação colaborativa é uma prática que permitir que qualquer usuário faça anotações nos conteúdos postados de forma livre e com qualquer tipo de marcas, é uma forma comum de organizar o conteúdo para uma navegação futura, filtragem ou pesquisa [Golder e Huberman 2006; Kelkar, John e Seligmann 2007], desempenhando um papel fundamental no compartilhamento de conteúdo em redes sociais [John e Seligmann 2007].

Esta marcação feita de forma colaborativa é descrito como "folksonomia". O termo é um neologismo de duas palavras, “folk” e “taxonomia”, que descreve estruturas conceituais criadas por usuários [Mika 2005; Schmitz 2006], criado por Thomas Vander Wal é uma analogia ao termo taxonomia tendo como principal característica a criação de tags (descritores) a partir do linguajar das pessoas que a utiliza [Rufino 2010].

Ao contrário de taxonomia que é comumente usado para organizar os recursos para formar uma hierarquia de categorias, folksonomia é não-hierárquica e não-exclusiva [Lambiotte e Ausloos 2006]. Por não ser hierárquico este sistema de marcação permite o uso de uma variedade de termos simultaneamente: *tags* gerais e as específicas [Ji et al 2007].

A atribuição de etiquetas é feita num ambiente social (compartilhado e aberto a outros) [Catarino e Baptista 2007] e descreve o universo de marcação colaborativa e esforços de indexação sociais que surgem em vários ecossistemas da Web [Russell 2013]. Os usuários com interesses comuns tendem a usar conjunto semelhante de marcas na marcação de recursos de interesse [Guo e Joshi 2010]. A semelhança dos usuários pode ser calculada através do padrão de marcação para recomendar propósito, pois proporciona relações sociais entre os usuários [Mika 2007; Siersdorfer e Sizov 2009]. Isso pode permitir uma maior colaboração entre usuários que compartilham interesses semelhantes.

2.2 Maioridade Penal no Brasil

A maioridade penal atualmente é um tema contemporâneo e bastante polêmico entre os legisladores, juristas e brasileiros em geral, assunto esse que congregam múltiplos olhares [Rocha 2015]. O autor afirma ainda que:

> Os meios de comunicação em geral revelam uma lógica conflitante de ordem social, e nesse cenário a população brasileira se divide entre aqueles que apoiam para que haja a redução da maioridade penal e aqueles que têm um posicionamento contrário a essa opinião. Surgem debates em todas as esferas do poder [Rocha 2015].

Segundo Oliveira (2015) o debate acerca da redução do patamar da responsabilidade penal juvenil circunda a sociedade brasileira, dividindo-a entre aqueles que acreditam que os adolescentes infratores cometem crimes porquanto não são suficientemente punidos e os que defendem ser uma medida de cunho apenas eleitoreiro que ataca somente o sintoma, mas não a causa do problema.

Mesmo já tendo a Câmara dos Deputados aprovado a Proposta de Emenda à Constituição (PEC) que reduz de 18 para 16 anos a idade penal para crimes hediondos, homicídio doloso e lesão corporal seguida de morte, menos de 24 horas após o plenário rejeitar a redução da maioridade para crimes graves, os debates nas redes sociais continuam, mostrando que o assunto ainda não foi finalizado.

3. MATERIAL E MÉTODOS

Em estruturas complexas como as que emergem das redes sociais pode-se utilizar o processo KDT (Knowledge Discovery from Text) na busca por padrões, tendências e regularidades em textos escritos em linguagem natural. Segundo [Wives 2002] a mineração de textos pode ser entendida como a aplicação de técnicas de KDD (Knowledge Discovery in Database) sobre dados extraídos de textos. Entretanto, KDT não inclui somente a aplicação das técnicas tradicionais de KDD, mas também qualquer técnica nova ou antiga que possa ser aplicada no sentido de encontrar conhecimento em qualquer tipo de texto [Moura 2004].

Para descoberta de conhecimento a partir de informações textuais, como as disponíveis no *Twitter* se dividiu a mineração em etapas bem características em um fluxo procedimental conforme foi criado por [Aranha 2007]. Todo o processo de mineração foi realizado utilizando-se as funções da linguagem de programação estatística R. A mineração de textos se contruiu através de quatro etapas: i) coleta dos dados; ii) pré-processamento dos dados coletados; iii) indexação; e iv) mineração e análise dos dados.

Os dados utilizados na identificação das possibilidades sobre a mineração de dados através do R foram coletados na rede social *Twitter*, por meio *hashtag* #MaioridadePenal. A consulta foi executada no dia 08 de abril de 2015, data da instalação da comissão que responsável por analisar a Proposta de Emenda Constitucional (PEC) para redução da maioridade penal, onde foram extraídos 1.755 tweets e re-tweets. Para se extrair dados do *Twitter* através do R alguns passos e funções foram necessários como mostra a Figura 1.

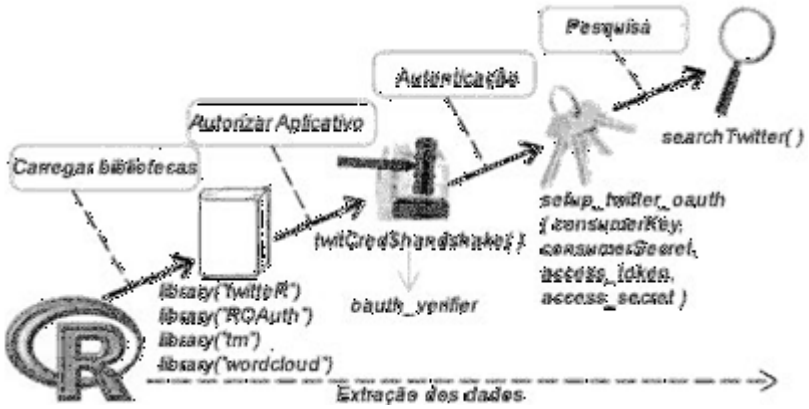

Figura 1: Passos e funções para extração de dados do Twitter através do R.

Tendo coletado os dados fez-se necessário um pré-processamento. Nesta etapa existe uma grande quantidade de informações contidas nos dados, que não são particularmente úteis para a mineração de texto, e precisam ser descartadas, tais como simbolos de pontuação, espaços em branco, números e palavras que são muito frequentes como as stopword. Stopword é uma palavra que não carrega significado na linguagem natural e portanto podem ser descartadas [Borges e Mourão 2013].

Há disponivel para R um pacote com inumeras funções para mineração de texto chamado "tm". Neste pacote a estrutura principal para o gerenciamento de documentos é chamado *Corpus*, que representa um conjunto de documentos de textos. Segundo [Feinerer 2014] o *corpus* é um conceito abstrato, e não pode existir várias implementações em paralelo. A implementação padrão é o chamado Vcorpus (Abreviação de Volátil Corpus) e são objetos do R mantidos totalmente na memória. Após o pré-processamento a estrutura de dados será indexada a partir dos termos existentes Corpus (Figura 2). Este procedimento tornará possível a recuperação rápida do conteúdo. Para isso o Corpus foi transformado em uma matriz de termos em documentos (TDM).

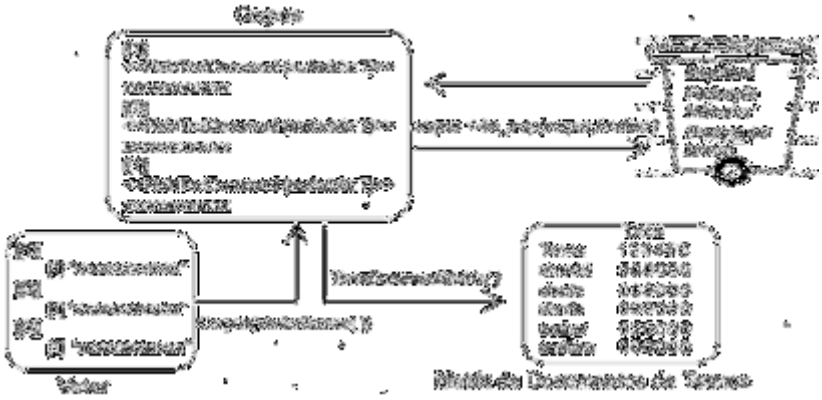

Figura 2: Etapas do pré-processamento e indexação dos dados.

Após a indexação, os dados já estão em formatos combatives para a geração de mapeamentos gráficos, seja por técnicas de clusterização ou de cloud word (nuvem de palavras).

As palavras em nuvem é uma trama comumente usada para visualizar um discurso ou conjunto de documentos de uma forma sucinta. Segundo [Stats 2012] este tipo de infográfico é muito utilizado na web por apresentar a informação quantitativa de maneira equilibrada e com um bom design o que mantem o leitor interessado.

No R o pacote "*wordcloud*" possibilita a criação de gráficos em formato de nuvem de palavras. O primeiro mapeamento gráfico foi feito da matriz de termos em documentos (TDM) com palavras com frequencia mínima no valor 10.

Figura 3: Nuvem de palavras gerado no R a partir da *hashtag* #MaioridadePenal

Outra maneira de obter um sentido visual é através da técnica de clusterização. Agrupamento ou *clustering* é uma forma de encontrar associações entre itens [Danneman e Heimann 2014], isto é, o processo de agrupar conteúdos baseados na informação difusa, como palavras ou frases em um conjunto de documentos [Fung 2001].

A função *hclust*( ) em R permite realizar agrupamentos, do tipo hierárquico aglomerativo. Segundo [Tan et.al 2006] esta abordagem de agrupamento refere-se a um conjunto de técnicas que inicia-se com cada ponto como um cluster individual. A cada passo une o par de pontos mais próximos até que reste somente um ou k *clusters*. Para geração do *cluster* foi passado como parâmetro a função o valor de correlação na matriz termo-documento o valor de 0.8 (Figura 4). Vale ressaltar que o valor pode variar de 0 a 1 de acordo com a correlação.

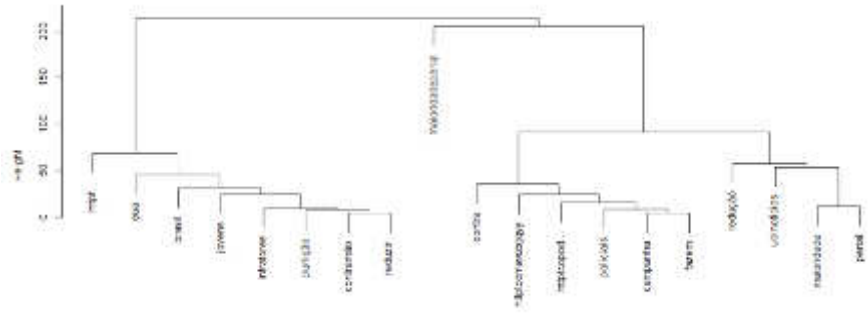

Figura 4: *Cluster* gerado no R a partir da *hashtag* #MaioridadePenal.

4. RESULTADOS E DISCUSSÃO

Através da mineração dos comentários postados na rede *Twitter* no dia 08 de abril de 2015 acerca da redução da maioridade penal usando-se o R, pode-se realizar três inferências. A primeira destaca uma campanha de policiais contra a redução da maioridade penal. Este resultado se deu pelos destaques na nuvem de palavras (Figura 2), pelo grupo criado na clusterização (Figura 3) e pelo resultado obtido por meio da função *findAssocs*( ) que retornou a correlação da palavras associadas ao termo "contra".

A segunda inferência foi a respeito do canal de noticias da UOL, que ocupou um local estratégico no debate sobre a redução, porém não se identificou se o posicionamento era a favor ou contrário, podendo ser isto demonstração de imparcialidade. O canal teve destaque na nuvem e apareceu no terceiro grupo formado a partir da clusterização. Interpretou-se como um veículo

que se posicionou estrategicamente através da *hashtag*, onde os tweets postados foram retransmitidos em uma quantidade superior aos demais usuários da rede. A participação destacável deste veículo se confirmou ao se totalizar o número de tweets e retweets coletados onde a fonte de origem foi o usuário @uolnoticias. O valor correspondeu a 24,61% do total coletados.

A terceira inferência identificou dois atores (usuários) que se destacaram com posicionamentos contrários sobre o Brasil e o Estados Unidos estarem ou não em contramão sobre o tema discutido, um foi o canal de notícias BBC Brasil, e o outro foi o deputado federal pelo Rio de Janeiro, Alessandro Molon. Os posicionamentos contrários dos dois tweets geraram a clusterização mostrada no primeiro cluster (Figura 3).

Os tweets postados por estes usuários foram: 1) #MaioridadePenal Na contramão do Brasil, EUA reduzem punição a jovens infratores; e 2)Nos EUA, estados que reduziram a #maioridadepenal agora querem aumentá-la, mostra matéria da The Economist.

5. CONCLUSÕES

Ao realizar a análise do contexto social da *hashtag* #MaioridadePenal fica claro a explosão de conteúdo carregado de sentimento que está disponível e que as funções disponíveis no R tornam a mineração de dados social um campo acadêmico próspero e com um domínio comercial crucial. A mineração de dados a partir de folksonomias e usando-se a ferramenta R demonstrou ser uma maneira poderosa para resumir redes e identificar pessoas-chave ou objetos que ocupam locais estratégicos e posições dentro da matriz de relações, como foi demonstrado e descrito neste estudo.

Conforme afirmou [Russell 2013] a partir de um tema arbitrário de interesse, o poder do *Twitter* e o conhecimento que se pode ganhar minerando seus dados tornam-se muito mais evidente.

A linguagem R apresentou-se muito poderosa, uma suite simples e extremamente extensível. Por possuir uma capacidade de instalação de pacotes diversificada fornecidos por uma comunidade acadêmica e profissional robusta permitiu a execução de todas as etapas da mineração proposta com facilidade.

A respeito da rede social *Twitter* evidenciou-se através da mineração que a interação, por meio da conexão, acontece com qualquer pessoa no mundo, sejam elas conhecidas ou não. Um tema ou tópico como a redução da maioridade penal pode ser discutido, unindo ideias em torno de valores e interesses de forma autônoma conforme afirmou [Marteleto 2001].

Quanto à mineração de dados sociais demonstra ser uma prática importante para a análise e descoberta de tendências e fatos importantes, sendo um procedimento ágil para descoberta de conhecimento quando se faz uso de ferramentas como o R. Assim como afirma Amorim (2006, p. 10): “Uma excelente prática de administração de dados é o enriquecimento dos dados, gerando ainda mais informação e conhecimento, melhorando assim a sua qualidade”. Dessa forma a mineração de dados sociais pode realizar varias descobertas importantes, seja para a empresa geradora dos dados ou para os usuários do sistema. A mineração de dados torna-se, então, uma das maiores fontes de investigação de informações e que além de ajudar na presente pesquisa, pode também ajudar em outras em diferentes aspectos.

REFERÊNCIAS

Amorim, Thiago. (2006) **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados** Disponível em: http://www.cin.ufpe.br/~tg/2006-2/tmas.pdf Acesso em: 09/07/2015 pag. 10.

Borges, L. e Mourão, L. (2013) **Recuperação de Informação - Conceitos e Tecnologia das Máquinas de Busca** 2 ed, Editora Bookman, pag.207.

Castells, M. (2003) **A galáxia da internet: reflexões sobre a internet, os negócios e a sociedade**, Tradução Borges, Maria Luiza X. De A. Editora Zahar 1ed.

Catarino, M.E e Baptista, A. A. (2007) **Folksonomia: um novo conceito para a organização dos recursos digitais na Web**, DataGramaZero - Revista de Ciência da Informação - v.8 n.3 , http://www.dgz.org.br/jun07/Art_04.htm

Danneman, N. e Heimann, R. (2014) **Social Media Mining with R** Editora Packt Publishing Ltd. 1 ed. , ISBN 978-1-78328-177-0

Feinerer, I. (2014) **Introduction to the tm Package - Text Mining in R**, http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf , Abril

Ferreira, G. C. (2011) **Redes Sociais de Informação: uma história e um estudo de caso** Perspectivas em Ciência da Informação, v.16, n.3, p.208-231

França, G. (2009) **Os ambientes de aprendizagem na época da hipermídia e da educação a distância**, Perspectivas em Ciência da Informação, v.14,n.1, p. 55-66, http://www.scielo.br/pdf/pci/v14n1/v14n1a05.pdf

Fung, G. (2001) **A Comprehensive Overview of Basic Clustering Algorithms** http://pages.cs.wisc.edu/~gfung/clustering.pdf

Golder, S. A., & Huberman, B. A. (2006). **Usage patterns of collaborative tagging systems. Journal of information science**, 32(2), 198-208.

Guo, Y., & Joshi, J. B. (2010, June). **Topic-based personalized recommendation for collaborative tagging system. In Proceedings of the 21st ACM conference on Hypertext and hypermedia** (pp. 61-66). ACM.

Hansen, D. L., Shneiderman, B. e Smithanalyzing, M. A. (2011) **Social media networks with NodeXL insights from a connected world**, Morgan Kaufmann Elsevier Inc. ISBN : 978-0-12-382229-1

Ji, A. T., Yeon, C., Kim, H. N., & Jo, G. S. (2007). **Collaborative tagging in recommender systems.** AI 2007: Advances in Artificial Intelligence, 377-386.

Kelkar, S., John, A., & Seligmann, D. D. (2007, March). **An Activity-based Perspective of Collaborative Tagging.** In ICWSM.

Lambiotte, R., & Ausloos, M. (2006). **Collaborative tagging as a tripartite network.** In Computational Science–ICCS 2006 (pp. 1114-1117). Springer Berlin Heidelberg.

Lévy, P. (2001) **A conexão planetária: o mercado, o ciberespaço, a consciência**. São Paulo: Editora 34.

Marteleto, R. M (2001) **Análise de redes sociais- aplicação nos estudos de transferência da informação** Ci. Inf., Brasília, v. 30, n. 1, p. 71-81. http://www.scielo.br/pdf/ci/v30n1/a09v30n1.pdf

Mika, P. (2005). **Ontologies are us: A unified model of social networks and semantics.** In The Semantic Web–ISWC 2005 (pp. 522-536). Springer Berlin Heidelberg.

Moura, M. F. (2004) **Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos**, Embrapa Informática Agropecuária, ISSN 1677-9274, 2004.

Oliveira, E. G. D. (2015). **Redução da maioridade penal: manobra eleitoreira ou ação eficaz contra a criminalidade e a violência?**

Poloni, K. e Tomaél, M. I. (2014) **Coleta de dados em plataformas de redes sociais: estudo de aplicativos** III Workshop de Pesquisa em Ciência da Informação (WPCI) pág. Web. 13 Abr. 2015

Recueiro, R. (2014) **Contribuições da Análise de Redes Sociais para o estudo das redes sociais na Internet: o caso da hashtag #Tamojuntodilma e #CalaabocaDilma**, Revista Fronteiras, doi: 10.4013, http://raquelrecuero.com/fronteirasrecuero2014.pdf

Rocha, S. B. da (2015). **A redução da maioridade penal**. Portal Âmbito Jurídico. Disponível em: < http://ambito-juridico.com.br/site/?n_link=revista_artigos_leitura&artigo_id=13332&revista_caderno=12 >. Acesso em 29 jul 2015.

Rufino, A. (2010) **Folksonomia: novos desafios do profissional da informação frente às novas possibilidades de organização de conteúdos** XXXII ENEBD – Encontro Nacional ee Estudantes de Biblioteconomia, Documentação, Gestão e Ciência da Informação - Múltiplos Olhares em Ciência da Informação, v.1, n.1.

Schmitz, C., Hotho, A., Jäschke, R., & Stumme, G. (2006). **Mining association rules in folksonomies**. In Data Science and Classification (pp. 261-270). Springer Berlin Heidelberg.

Siersdorfer, S., & Sizov, S. (2009, June). **Social recommender systems for web 2.0 folksonomies.** In Proceedings of the 20th ACM conference on Hypertext and hypermedia (pp. 261-270). ACM.

Stats, F. (2012) **Words in politics: some extensions of the word cloud** http://blog.fellstat.com/?p=101

Tan, P.N., Steinbach, P., Kumar,V. (2006) **Introduction to Data Mining** http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Wal, T. V. (2005). **Folksonomy definition and Wikipedia**. Bethesda, Maryland. Disponível em: http://www.vanderwal.net/random/entrysel.php?blog=1750> Acesso em 10 de abr 2015.

Witten, I. H., Frank E. e Hall M. A. (2011) **Data Mining Practical Machine Learning Tools and Techniques.** 3. Ed. Morgan Kaufmann, Elsevier, ISBN 978-0-12-374856

WIVES, K. L. (2002) **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva** Exame de Qualificação EQ-069, PPGC-UFRGS. http://www.leandro.wives.nom.br/pt-br/publicacoes/eq.pdf