UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Nuvens de tags no Twitter: estudo e implementação

Luís Filipe Auto Gomes

Trabalho de Graduação

Monografia apresentada ao Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do Grau em Ciência da Computação.

Orientador: Ricardo Bastos Prudêncio

RECIFE, JULHO DE 2011

“A sabedoria começa na reflexão”.
Sócrates

Agradecimentos

Registro aqui meus sinceros agradecimentos a Deus e à minha família, por tudo, aos meus amigos de turma, que partilharam comigo de todos momentos deste curso e ao meu orientador, pelo seu apoio, compreensão e, sobretudo, pela simpatia sincera.

Resumo

O Twitter pode ser visto como uma vasta fonte de informações de todos os tipos. Trata-se de uma base de conhecimento que pode ser usada para a obtenção de opiniões, idéias, fatos e sentimentos. É uma nova plataforma para a disseminação de informação e que, a cada dia, vem sendo utilizada por várias pessoas e para diversos fins. As nuvens de tags constituem uma técnica para visualização e análise de informações em páginas da web. Trata-se de uma abordagem que permite uma compreensão muito mais rápida do contexto associado a um documento web, de maneira que são objetivos deste trabalho: fazer uma análise relativa ao processo de criação de uma nuvem de tags e implementar um sistema que permita a visualização de uma nuvem de tags baseada no resultado de uma consulta qualquer às mensagens do Twitter.

Palavras-chave: Twitter, Nuvens de tags, Tomada de decisão, Opinião na Internet.

Abstract

Twitter can be seen as a vast source of information of all kinds. It is a knowledge base that can be used to obtain opinions, ideas, facts and feelings. It is a new platform to spread information that is being used by several people, for different purposes. The use of Tag clouds is a technique for viewing and analyzing information on web pages. It is an approach that allows a much more rapid understanding of the context within a web document. The main goal of this study is to analyze the process of creating a tag cloud and to implement a system that allows the visualization of a tag cloud based on the result of a query on twitter's database.

Keywords: Twitter, Tag Clouds, Decision-making Process, Internet Reviews.

Sumário

1 – Introdução

O objetivo desse trabalho é explanar como a técnica de visualização de nuvens de tags pode ser aplicada sobre consultas genéricas à base de dados do Twitter.

O estudo será iniciado Capítulo 2, com uma análise do contexto inerente ao trabalho, contemplando os aspectos gerais relativos ao Twitter e às nuvens de tags.

Após a contextualização, procede-se, no Capítulo 3, à revisão de alguns conceitos importantes da área de Recuperação de Informação, necessários à compreensão de certas etapas do processo de geração de uma nuvem de tags.

Em seguida, no Capítulo 4, as etapas para a construção de uma nuvem de tags relativa a mensagens do Twitter serão explicadas em detalhes, envolvendo aspectos conceituais e de implementação, desde à consulta às referidas mensagens até o processo de visualização da nuvem.

No Capítulo 5 são apresentados alguns experimentos realizados em um programa desenvolvido com base no processo apresentado neste trabalho.

Por fim, remete-se à conclusão e às referências bibliográficas.

2 – Visão geral: Twitter e Nuvens de tags

Antes do início do estudo, serão tratados, neste capítulo, de forma um pouco mais detalhada, os temas tidos por alicerces desse trabalho.

2.1 – Twitter

Atualmente, é notável a proliferação dos chamados weblogs no meio da Internet. Pode-se dizer que constituem um dos fenômenos com mais importância e visibilidade na seara das tecnologias Web 2.0 [8]. São meras páginas da internet, sujeitas a atualizações ao livre critério do administrador, de modo que os conteúdos pregressos são mantidos e organizados conforme uma ordem cronológica.

Nas referidas páginas as pessoas escrevem sobre o que bem entenderem, de maneira que hoje é possível encontrar na "blogoesfera" temas dos mais diversos, saindo do besteirol, passando pela utilidade pública e alcançando até a comunidade científica.

Em 2006, o caráter social da internet já era tamanho que, a análise de tráfego dos sites de redes sociais mostrava que essas estavam prestes a alcançar os grandes portais como Google e Yahoo [9]. Em parte, o crescimento desse tipo de mídia justifica-se pela composição da usabilidade, do aspecto colaborativo e, especialmente, da possibilidade de projeção da subjetividade do autor do blog em cada uma de suas postagens, as quais refletem seus sentimentos e opiniões [8].

Assim sendo, é pacífica a importância dos blogs dentro do contexto atual da internet. No entanto, nos anos mais recentes, é evidente o crescimento de uma outra tecnologia bastante peculiar: o microblog. Microblogging ocorre quando um usuário escreve, em vez de um texto longo, atrelado a fotos e vídeos, como ocorre em um blog ordinário, uma pequena mensagem, a qual traz informações breves sobre a vida do autor, um pensamento fortuito, um desejo momentâneo. Via de regra, tais mensagens não superam 200 caracteres [7]. Uma vez escrito, o pequeno texto pode ser enviado para uma rede, na qual vários outros usuários terão a chance de lê-lo, bem como também postar as

suas próprias mensagens. Entre os serviços atuais de microblogging destacam-se o Jaiku, Pownce e, em especial, o Twitter [7].

O Twitter é uma rede social específica para microblogging. Em suma, é uma comunidade por meio da qual os usuários enviam e recebem mensagens curtas, de até 140 caracteres. A idéia inicial era simplesmente informar ao mundo o que o usuário estava fazendo no momento do envio da mensagem: no que estava pensando, qual era a música que estava ouvindo, se estava prestes a sair, se havia chegado de algum lugar, etc. Com o tempo, esse cenário começou a se expandir. O Twitter, atualmente, pode ser visto como uma vasta fonte de informações de todos os tipos. Trata-se de uma base de conhecimento que pode ser usada para a obtenção de opiniões, idéias, fatos e sentimentos [10]. É uma nova plataforma para a disseminação de informação e que, a cada dia, vem sendo utilizada por várias pessoas e para diversos fins.

Hoje, encontram-se os usuários: velhos, jovens, políticos, empresários, lojas, jornais e revistas [10], de maneira que é possível ter acesso a informações relativas às promoções de uma loja, por exemplo, assim como é possível encontrar notícias sobre fatos que ocorrem no mundo inteiro.

Logo, é evidente que a natureza e o valor das informações mudaram bastante, assim como o volume das mesmas, de modo que é desejável a incorporação de técnicas eficientes para dar suporte ao processo de análise desse imenso e precioso repositório.

2.2 – Nuvens de tags

Tagging é uma abordagem para criação de meta dados, uma técnica cuja popularidade aumentou com o advento das chamadas ferramentas de bookmarking social [3], dentre as quais destacam-se o del.icio.us, Flicker e Technorati, sistemas em expansão acelerada no cenário web atual [6].

Por meio desses serviços, dentre outros sistemas de tagging, usuários podem adicionar as chamadas tags, meras palavras livremente escolhidas, a recursos específicos da web, com o intuito de classificá-los da forma que for conveniente para melhorar consultas posteriores aos respectivos recursos. Dessa forma, é possível afirmar que o processo de

Tagging existe, precipuamente, para fins de categorização. Não obstante, observando-se o referido processo sob uma ótica mais subjetivista, relativamente à possibilidade de indexação e construção de conhecimento, é evidente a sua conotação social [6].

O conjunto de dados oriundo do crescente processo de tagging na web é conhecido como ‘folksonomy', neologismo relativo às palavras “folk” e “taxonomy” [3]. Trata-se de um conjunto de palavras-chave, livres de qualquer relação semântica direta entre si. No entanto, quando as referidas palavras são analisadas por intermédio de modelos visuais específicos, há uma compreensão muito mais clara do contexto associado a um determinado conjunto de “folksonomy” [3]. Uma das abordagens é o uso das chamadas nuvens de tags ou tag-clouds.

As nuvens de tags são representações visuais de conjuntos de palavras, via de regra, escolhidos por meio de critérios específicos [5]. São os modelos visuais utilizados pela maioria dos serviços de bookmarking social , sendo que a "nuvem" é, na verdade, uma lista das tags mais populares de um conjunto de dados de 'folksonomy'. Na prática, a lista é composta pelas tags mais recorrentes do conjunto.

A função das nuvens é representar as variáveis de interesse, como a popularidade, por exemplo, associadas às respectivas palavras-chave por intermédio de elementos visuais como o tamanho, fonte e cor [2]. Em geral, as tags são exibidas em ordem alfabética, em fontes calculadas com base no peso de cada tag, de modo que as mais populares aparecem em fontes maiores e as menos populares em fontes menores. Também é comum a associação de cores às tags, como um meio para a atribuição de outros parâmetros de diferenciação.

Embora permitam a exibição de muito mais itens, de dezenas até centenas, via de regra, em termos conceituais, são similares a histogramas e gráficos de freqüência[4].

A depender do contexto, as nuvens de tags, segundo [5], podem ser úteis quando se deseja:

Pesquisar - Quando o usuário quer localizar termos específicos e importantes relativos a um determinado contexto.

Navegar - Quando o usuário usa as nuvens para navegar por páginas diversas, sem procurar por um tópico específico, no caso de haver associação das tags a hiperlinks.

Obter impressão - Quando o usuário, por intermédio da nuvem, consegue ter, rapidamente, uma impressão inicial do contexto associado às tags.

Reconhecer - Quando o usuário quer fazer a distinção de contextos associados a um mesmo termo.

Ainda, as nuvens de tags permitem uma compreensão muito mais rápida do contexto associado a um documento web [1], de modo que a intenção deste trabalho é aplicá-la às consultas relativas às mensagens do Twitter, com o objetivo encontrar as questões relevantes que estão circulando na rede em relação ao que foi consultado.

3 – Recuperação de Informação – conceitos importantes

Neste capítulo do trabalho serão discutidos alguns conceitos da área de Recuperação de Informação (RI) que são de grande relevância para o processo de geração de uma nuvem de tags.

3.1 – Visão geral do processo de RI

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.

O processo de recuperação de informação é composto por quatro etapas básicas: a preparação dos dados, a indexação, a pesquisa e a ordenação das informações recuperadas [11].

Os conhecimentos mais relevantes para este trabalho estão presentes no processo de preparação dos dados dos sistemas de RI. Na referida fase, o intuito principal é o de manipular e adicionar uma certa padronização às informações, criando o que se chama de visão lógica dos dados, que será, a depender da situação, a representação integral dos dados, ou uma versão bastante simplificada. No caso presente, as informações são os textos contidos nos tweets recuperados de uma consulta qualquer à base de dados do Twitter. Diferentemente dos sistemas comuns de RI, nos quais a visão lógica dos dados serve para aprimorar do processo de indexação [11], o que melhora as consultas posteriores aos respectivos dados, neste trabalho, a "visão lógica" constituirá, de certa forma, a própria nuvem de tags. O objetivo é fazer uma seleção automática dos termos mais freqüentes e relevantes presentes nos resultados das consultas.

Para a criação dessa representação simplificada das informações consultadas, é preciso recorrer a uma série de técnicas de operações sobre textos, comumente empregadas pelos sistemas de RI, as quais serão discutidas nas próximas páginas.

3.2 – Operações sobre o texto

Conforme comentado na seção anterior, é necessário que sejam realizadas algumas operações sobre os textos dos tweets consultados. O emprego dessas operações culminará na criação de uma representação simplificada dos respectivos tweets, uma representação que possuirá apenas os termos dotados de valor semântico relevante, o que facilitará a identificação dos termos mais importantes dentro do conjunto completo dos tweets retornados pela consulta do usuário.

A título de exemplo, se o texto de um dos tweets retornados pela consulta fosse:

"**Centro de Informática faz a transmissão simultânea do Startup Lessons Learned neste mês**"

O emprego das operações sobre o referido texto geraria, hipoteticamente, algo similar à seguinte lista de termos:

"**centro informática transmissão simultânea startup lessons learned mês**"

As operações básicas para esse tipo de manipulação textual, nos sistemas de RI, consubstanciam quatro etapas do processo de operações sobre o texto: a análise léxica, a remoção de stopwords, a fase de stemming e a identificação dos grupos nominais [11].

Na análise léxica, em suma, são removidos os caracteres indesejados presentes no texto, de maneira que os critérios adotados para a respectiva remoção, dentre outras alterações tratadas mais adiante, podem variar de acordo com o contexto de cada aplicação.

Na etapa de remoção de stopwords, 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.

Na fase de stemming, a idéia é reduzir os termos aos seus respectivos radicais. Trata-se de uma abordagem utilizada pelos sistemas de RI para aumentar a probabilidade

de *matching* dos termos indexados com as consultas dos usuários [11]. Em uma situação hipotética, na qual o usuário fizesse uma pesquisa utilizando o termo "programador", por exemplo, poderia ocorrer um *matching* com outros termos como "programa", "programação", "programar", etc.

Por fim, 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".

A seguir, serão tratadas as fases do processo de operações sobre o texto dos sistemas de RI mais significativas para a criação de uma nuvem de tags.

3.2.1 – Análise Léxica

A análise léxica vai encabeçar as operações sobre os textos, transformando-os em meras listas de termos individuais [11]. Para tal, serão removidos todos os espaços em branco, bem como todos os caracteres especiais, contemplando principalmente os sinais de pontuação. Também podem ser retirados os dígitos, uma vez que esses caracteres isolados, na maioria dos casos, possuem pouco valor semântico.

Quando há a intenção de se prezar um pouco mais pela padronização, também pode ser feita a conversão de todos os termos para o formato *lower case*.

Conforme explanado na seção anterior, na prática, a natureza da aplicação vai ditar os parâmetros para o processo de análise léxica. No caso presente, opta-se pela remoção de todos os espaços, caracteres especiais e dígitos dos tweets retornados pela consulta do usuário. Também será realizada a conversão de todas letras maiúsculas em letras minúsculas.

3.2.2 – Remoção de Stopwords

Stopwords são, em regra, termos de elevada recorrência nos mais diversos textos. São os pronomes, preposições, artigos, certos advérbios e verbos, dentre outros, de maneira que cada língua possui o seu conjunto de stopwords específico [11].

Essas palavras precisam ser retiradas da lista de termos porque, além de possuírem baixo teor semântico quando isoladas, o fato de estarem quase sempre presentes nos textos escritos em uma língua específica importa a adulteração das análises posteriores relativas à visão lógica dos respectivos textos.

A título de exemplo, sejam considerados os seguintes fragmentos de tweets:

"**F.V. volta às quadras de vôlei após quatro anos afastada**"
"**F.V. craque de bola, se desaposenta**."
"**F.V. de volta às quadras**"

Se a análise da freqüência dos termos levar em consideração o termo "de", este será classificado entre os mais importantes do conjunto, o que é indesejável, posto que a preposição, isoladamente, não possui qualquer relevância semântica relativa ao contexto apresentado pelos três textos. Projetando esse raciocínio para uma consulta que retorne cem tweets, é óbvio que as stopwords presentes, se consideradas, desvirtuariam a análise de freqüência.

No caso presente, para a remoção de stopwords, inicialmente são utilizadas duas listas com pronomes, verbos conjugados, artigos e preposições. Uma lista traz os termos para a língua portuguesa e a outra para a língua inglesa, de forma que as palavras constantes destas listas são eliminadas dos conjuntos de termos obtidos após a etapa de análise léxica.

4 – Construindo nuvens de tags a partir do Twitter

Nas próximas páginas será explanado o processo de geração e exibição de uma nuvem de tags relativa a uma consulta às mensagens do Twitter, conforme apresentado na figura abaixo.

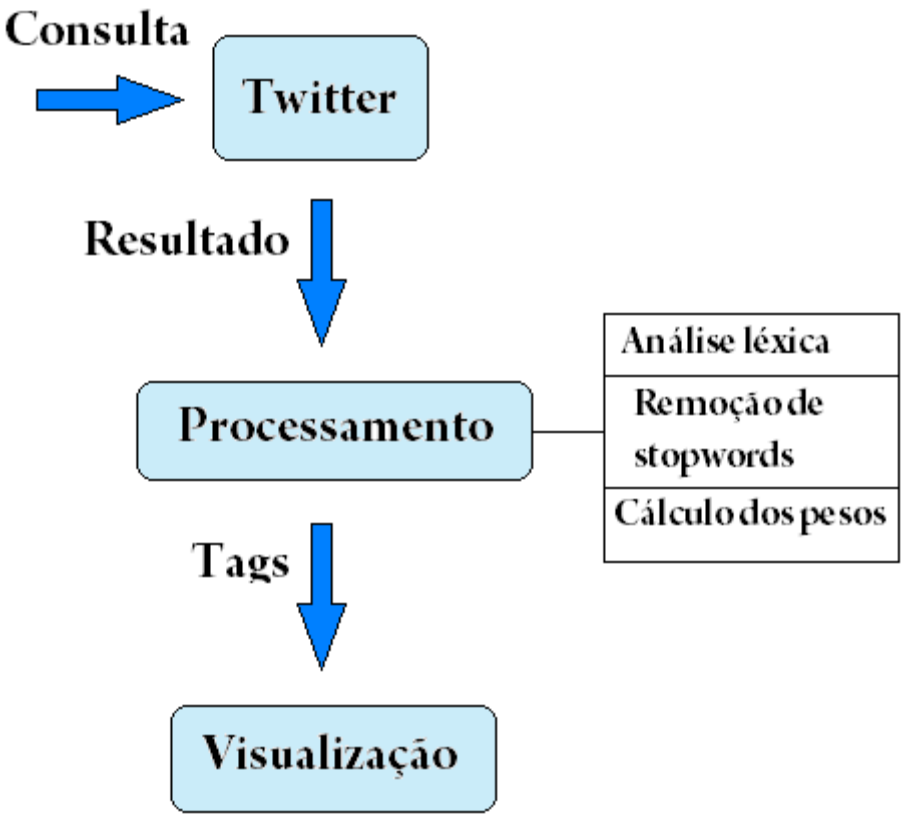

Figura 4.1 – visão global do processo de geração da nuvem de tags a partir do Twitter.

As tecnologias utilizadas e sugeridas por esse trabalho são: a linguagem de programação Java, a Interface Eclipse e a API Twitter4j.

4.1 – Consulta à base de dados do Twitter

A consulta é o ponto de partida, a etapa que irá trazer as informações necessárias para geração da nuvem.

A API Twitter4j fornece uma infra-estrutura razoável e relativamente simples para a realização de consultas à base de dados do Twitter.

Para criar uma consulta com o Twitter4j, é necessário instanciar um objeto do tipo Query, sendo que, no momento da respectiva instância já é possível atribuir um valor ao texto da consulta, conforme o exemplo abaixo:

Query consulta = new Query(“consulta do usuário”);

A string “consulta do usuário” já é o efetivo texto da consulta, não sendo necessário mais nenhum esforço para criar uma consulta simples usando a representação do Twitter4j. Entretanto, a API ainda apresenta algumas possibilidades de parametrização das consultas.

O método setRpp() do tipo Query, por exemplo, permite a definição do número de tweets retornados por página, já o método setLocale(), embora ainda não funcione perfeitamente, permite a restrição da língua dos respectivos tweets, dentre outros parâmetros, os quais podem ser estudados em sua totalidade na documentação do Twitter4j.

Os únicos parâmetros efetivamente definidos para os fins desse trabalho foram: o valor do texto da consulta, fornecido pela entrada do usuário, e o número de tweets por página, o qual foi estabelecido como o máximo (100).

Criada a consulta, é necessário executá-la para obter os resultados que servirão de base para a geração da nuvem. Essa execução deve ser precedida da abertura de uma sessão com o Twitter, o que pode ser feito conforme a seguinte linha de código:

Twitter twitter = new TwitterFactory().getInstance(usuario, senha);

Com a abertura da sessão, procede-se à execução da consulta, parametrizando-se um objeto do tipo Query no método search (Query query) da classe Twitter:

```
QueryResult resultado = twitter.search(consulta);
```

A classe QueryResult representa o resultado da consulta parametrizada e pertime, sem maiores dificuldades, o acesso à lista dos tweets retornados pelo método search.

Uma vez com o resultado da consulta em mãos, pode-se prosseguir para a etapa de processamento dos respectivos textos.

4.2 – Processamento dos Tweets

Conforme explicitado anteriormente, para a obtenção de uma representação lógica que contenha apenas os termos mais importantes de cada tweet, faz-se necessária a aplicação de uma série de operações sobre os textos presentes em cada um deles. Posto isso, uma vez com os tweets retornados do resultado da consulta à disposição, a implementação deve seguir um direcionamento no qual o objetivo é a "limpeza" e padronização desses tweets.

4.2.1 – Análise léxica

Para cada texto, deve ser feita inicialmente a análise léxica. Como visto, essa etapa do processamento irá retirar todos os dígitos e caracteres especiais. Também serão passadas todas as letras para o formato *lower case*. Esse tipo de processamento pode ser feito sem maiores problemas por intermédio dos recursos oferecidos pela classe String da linguagem Java.

Todavia, pelo fato de o Twitter ser uma rede global, contemplando mensagens escritas em várias línguas e expressas, em sua maioria, em linguagem informal, repletas de gírias e códigos específicos, alguns deles relativos ao contexto da própria rede, como é o caso do uso do "@" precedente dos nomes de usuários, do "#" para a definição das

hashtags, dentre outros, a etapa de análise léxica precisa ser pouco mais específica para a situação.

No caso em questão, por exemplo, optou-se pela manutenção do caracter "@", quando associado a um nome de usuário, dada a valoração semântica que o símbolo agrega ao respectivo nome.

Ainda, muitas mensagens possuem URLs em seus conteúdos, que também exigem identificação e tratamento especial, pois a mera remoção dos caracteres especiais pode gerar um resultado indesejado. Neste trabalho, para evitar esse tipo de problema, as URLs presentes nos textos são identificadas e removidas.

Outra peculiaridade nos tweets é o símbolo "RT", que representa o *Retweet* da respectiva mensagem. Basicamente, significa que o usuário está disseminando uma mensagem de outra pessoa. O "RT" é um dos símbolos quase sempre presentes nas mensagens do Twitter, não possui significado revelante para uma nuvem de tags e, no entanto, passa sem problemas pela análise léxica padrão, de maneira que a sua manutenção afetaria negativamente o resultado final. Posto isso, opta-se pela remoção desses símbolos também na etapa da análise léxica.

4.2.2 – Remoção de stopwords

Uma vez com os textos preliminarmente processados, procede-se à etapa de remoção de stopwords. Para essa etapa, é mantido em disco um arquivo com duas listas de stopwords, uma para a língua portuguesa e outra para a língua inglesa.

Cada termo resultante do primeiro processamento é comparado com as palavras das listas, de maneira que, havendo o casamento com alguma dessas palavras, o termo é removido da representação do tweet. Para fins de otimização, à medida em que as stopwords são identificadas, elas são escritas em memória para posterior comparação sem a necessidade de uma nova leitura de todo o arquivo em disco.

As peculiaridades das mensagens do Twitter exigem um cuidado extra também nessa etapa de remoção de stopwords. A mera consideração de stopwords conforme a norma culta padrão pode gerar resultados indesejados, uma vez que é recorrente o uso de

abreviações nas mensagens. A título de exemplo, o pronome "você" é largamente escrito como "vc", a preposição "para" como "p/", o verbo "está" como "tá", "you" como "u", "why" como "y", etc.

É preciso considerar uma lista extra de stopwords que contemple essas particularidades, afim de minimizar os efeitos negativos que o linguajar eminentemente informal dos tweets pode causar no resultado final da nuvem.

4.2.3 – Cálculo dos pesos

Uma vez com apenas os termos desejados em mãos, pode-se começar o cálculo das respectivas freqüências, as quais irão influenciar diretamente o resultado final da nuvem.

É comum a utilização da técnica TF-IDF para o cálculo de pesos nos sistemas de RI [1]. No entanto, no trabalho presente foi feita a opção por um cálculo simplificado de freqüências, o qual leva em consideração, pura e simplesmente, quantas vezes um determinado termo recorre no conjunto completo.

É realizado ainda um tratamento especial relativo aos termos próprios da consulta do usuário. O fato é que esses termos tendem a aparecer em todo o resultado da consulta, de maneira que seriam classificados, quase sempre, como os mais importantes da nuvem de tags e, portanto, seriam exibidos em fontes muito grandes. Para evitar esse problema, os termos da consulta do usuário são desconsiderados para efeito do cálculo de freqüências.

4.3 – Visualização

Com os termos e seus respectivos pesos à disposição, pode-se partir para a geração da nuvem propriamente dita. O primeiro passo é fazer o mapeamento dos pesos em fontes de texto que possam representar os respectivos termos adequadamente.

Para isso, é recomendada a organização dos valores de freqüências em grupos de importância, o que pode ser feito por meio do seguinte cálculo, conforme demonstra [1]:

$$Grupo(t) = \left\lfloor \left(\frac{t - wmin}{wmax - wmin} \right)^{\beta} * k \right\rfloor$$

Para cada termo, o cálculo indicará o grupo ao qual ele pertence, de modo que os termos desse mesmo grupo serão representados pela mesma fonte. A variável k representa o número de grupos com o qual se deseja trabalhar. O valor da freqüência do termo é representado por t, as variáveis $wmax$ e $wmin$ representam, respectivamente, o menor e o maior dos valores de freqüência do conjunto total de termos e o parâmetro β altera a distribuição dos termos nos grupos de importância, de maneira que, quanto maior for o seu valor, menor será a proporção entre as fontes dos termos mais importantes.

Identificado o grupo de cada termo, estes podem ser exibidos na fonte característica do respectivo grupo. Para tal, sugere-se o uso de tags html do tipo em conjunção com o tipo JPanel da linguagem Java.

Em cada tag é parametrizada a fonte do termo, conforme o exemplo abaixo:

**<span style = “font-size: [FONTE];”> [TAG] **

Definido o texto html completo, basta parametrizá-lo na instância de um objeto do tipo JLabel, que pode ser adicionado em qualquer painel da infra-estrutura de interface da linguagem Java (Java.Swing).

O resultado da execução da aplicação será a exibição da nuvem de tags da consulta realizada pelo usuário.

5 – Experimentos e resultados

Esse capítulo presta-se à apresentação de nuvens de tags geradas por um programa desenvolvido especialmente para o trabalho presente.

A consulta “cin ufpe” foi realizada várias vezes com o referido programa, sendo que alguns dos parâmetros explicados no capítulo anterior são definidos de formas diferentes, afim de esclarecer os efeitos que a alteração de cada um deles produz no resultado final da nuvem de tags.

5.1 – Quantidade de grupos

Nessa seção, será alterado o valor que representa a quantidade de grupos na qual será distribuída a nuvem de tags. Trata-se do valor k, que serve também como referência para determinar as fontes das tags exibidas (ver seção 4.3).

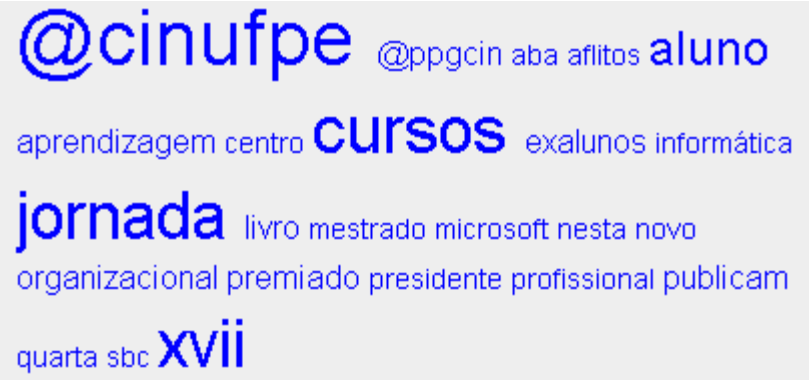

Figura 5.1 – primeira consulta para análise dos efeitos da quantidade de grupos k. Os valores utilizados como parâmetros são: $\mathbf{k = 50}$ e $\beta = 1.2$.

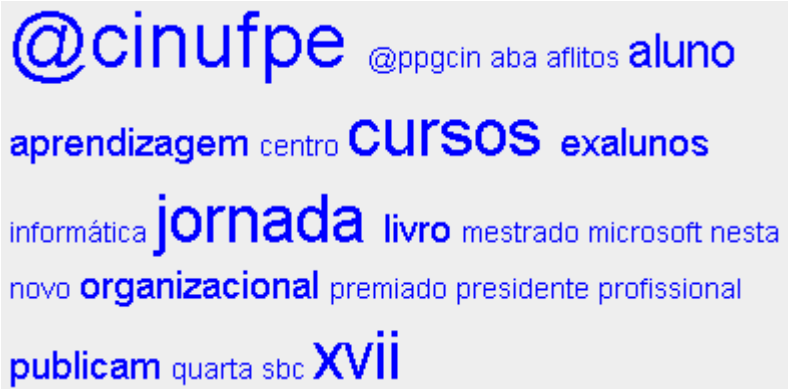

Figura 5.2 – segunda consulta para análise dos efeitos da quantidade de grupos k. Os valores utilizados como parâmetros são: $\mathbf{k = 200}$ e $\beta = 1.2$.

Percebe-se uma diferença grande com relação às fontes presentes nas figuras 5.3 e 5.4.

O aumento do valor k em quatro vezes traz um número maior de grupos de importância para enquadrar as tags. Dessa forma, como na implementação presente os grupos dos termos determinam as suas fontes, as tags da figura 5.4 possuem fontes maiores e mais diversificadas do que as da figura 5.3.

5.2 – Distribuição nas grupos

Conforme visto anteriormente, a distribuição dos pesos em grupos de importância pode ser alterada por meio do parâmetro β (ver seção 4.3).

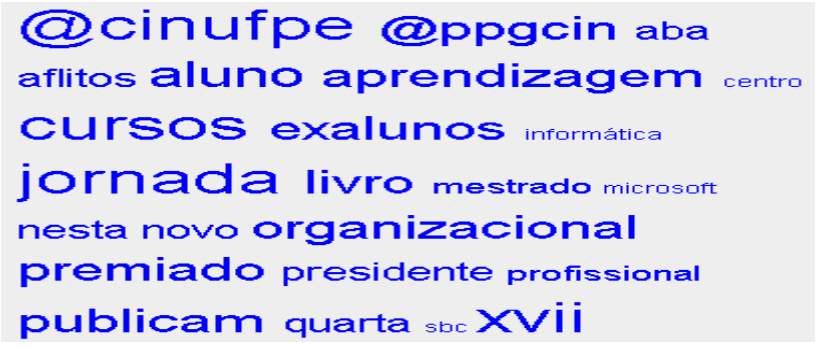

Figura 5.3 – primeira consulta para análise dos efeitos do parâmetro de distribuição nos grupos β. Os valores utilizados como parâmetros são: $k = 100$ e $\mathbf{\beta = 0.3}$.

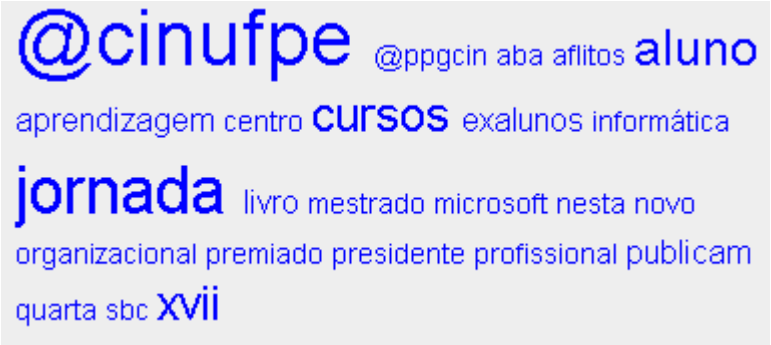

Figura 5.4 – segunda consulta para análise dos efeitos do parâmetro de distribuição nos grupos β. Os valores utilizados como parâmetros são: k = 100 e β **= 2.0.**.

É possível notar que a figura 5.6 possui uma diversidade maior em relação às fontes das tags. Além disso, tags exibidas com fontes iguais na figura 5.5 aparecem em fontes diferentes na figura 5.6, que é justamente o efeito que se obtém aumentando o valor do parâmetro β (ver seção 4.3).

5.3 – Parâmetros sugeridos

É evidente que as disparidades entre os resultados de diferentes consultas podem exigir definições diversas dos parâmetros citados nas seções anteriores. No entanto, segue uma configuração que apresenta resultados razoáveis para a maioria das consultas:

Figura 5.5 – consulta com os valores sugeridos. Os valores utilizados são: k **= 100 e** β **= 1.2**.

6 – Conclusão

Neste trabalho foram apresentados os conceitos básicos para a criação de uma nuvem de tags. Para tal, fez-se necessária uma análise breve de alguns princípios da área de Recuperação de Informação. Além disso, foram tratados aspectos técnicos específicos relativos à construção de nuvens de tags.

No entanto, o objetivo principal do estudo vai mais além: trata-se de uma análise do que pode resultar da aplicação dessa técnica de visualização sobre bases de dados ricas em informações heterogêneas, como é o caso do Twitter, conforme foi explanado nas seções iniciais.

Foi desenvolvido um programa em Java especialmente para realização de alguns experimentos técnicos com relação às nuvens de tags geradas a partir da base de dados do Twitter, de modo que a análise dos resultados revela a enorme gama de configurações diferentes que pode ser obtida mediante a alteração de parâmetros específicos da nuvem.

A idéia é que esse estudo sirva como base para quem desejar explorar a aplicação técnica das nuvens de tags sobre bases eminentemente heterogêneas, trabalhando em cima dos diversos parâmetros possíveis.

Referências Bibliográficas

[1] GOTTRON, T. **Document word clouds**: visualizing web documents as tag clouds to aid users in relevance decisions. Institut für Informatik, Johannes Gutenberg-Universität Mainz, Mainz, Germany, 2009.

[2] BATEMAN, S.; GUTWIN, C.; NACENTA, M. **Seeing things in the clouds**: the effect of visual features on tag cloud selections. University of Saskatchewan, Saskatoon, SK, Canada, 2008.

[3] HASSAN-MONTERO, Y.; HERRERO-SOLANA, V. Improving Tag-Clouds as Visual Information Retrieval Interfaces. Scimago Research Group. University of Granada, Faculty of Library and Information Science, Granada, SPAIN, 2006.

[4] KASER, O; LEMIRE, D. **Tag-Cloud Drawing**: Algorithms for Cloud Visualization. University of New Brunswick Saint John, NB, Canada; Université du Québec à Montréal ,Montréal, QC, Canada, 2007.

[5] RIVADENEIRA, A.W, et al. **Getting Our Head in the Clouds**: Toward Evaluation Studies of Tagclouds. Department of Psychology University of Maryland; Collaborative User Experience IBM Research. Cambridge, MA, USA, 2007.

[6] SINCLAIR, J.; CARDEW-HALL, M. **The folksonomy tag cloud**: when is it useful? Department of Engineering, The Australian National University, Canberra, Australia, 2007.

[7] JAVA, A., et al. **Why We Twitter**: Understanding Microblogging Usage and Communities. University of Maryland Baltimore County, Baltimore, USA; NEC Laboratories America, CA, USA, 2007.

[8] EBNER, M; SCHIEFNER, M. **Microblogging** – more than fun? Graz University of Technology, Graz, Austria; University of Zurich, Zurich, Switzerland, 2008.

[9] Social Networks gaining on Internet portals – Online, acesso em 02/06/2011 na url http://www.readwriteweb.com/archives/social_networks_vs_portals.php

[10] CHEONG, M.; LEE, V. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. Monash University, Melbourne, Australia, 2009.

[11] MANNING, D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to Information Retrieval. Cambridge University Press. 2008.