Folksonomias: Identificação de Padrões na Seleção de Tags para Descrever Conteúdos

Cleber Gouvêa[1], Stanley Loh[1,2]

[1]Catholic University of Pelotas (UCPEL)
[2]Lutheran University of Brasil (ULBRA)

cleber AT sindiq.com.br , sloh AT terra.com.br

Resumo

A categorização e organização do conteúdo através de tags (Folksonomia) têm se tornado popular na web como resultado ao surgimento de sites que privilegiam a colaboração. Trabalhos recentes têm demonstrado que as tags são úteis para categorizar informações de forma geral, sendo no entanto menos eficientes na indicação de seu conteúdo específico. Para suprir essa limitação, métodos automáticos de identificação de tags têm surgido. O presente trabalho busca aprimorar essas técnicas através da análise aprofundada de padrões de ocorrência de tags em um corpus de centenas de notícias cadastradas no site Delicious. Foram feitas estatísticas variadas relacionadas à posição da tag e sua freqüência dentro de cada texto. Busca-se assim compreender e auxiliar a construção coletiva do conhecimento em sistemas de Folksonomia, aprimorando a categorização e agrupamento de documentos e conseqüentemente aumentando a precisão para as consultas.
Palavras-chave: Folksonomia, Tags, Notícias, Inteligência Coletiva, Web 2.0.

Abstract

The categorization and organization of the content through tags (Folksonomy) became popular in the web with recent sites that utilize collaboration. Recent works have demonstrated that tags are useful to classify information in a general way, being however less efficient for indicating specific content. For managing that limitation, automatic methods for tag identification have been developed. The present work intends to improve those techniques through the discovery of patterns in tag frequency, analyzing a corpus of hundreds of news published in the site Delicious. The work presents statistics related to the position of the tags and their frequency inside the texts. The main goal is to understand and help the collective construction of knowledge in systems that utilize Folksonomies, improving the categorization and clustering of documents and consequently increasing the precision of the searches.
Key-words: Folksonomy, Tags, News, Collective Intelligence, Web 2.0.

1 Introdução

Folksonomia (em inglês, *folksonomy*) é um neologismo utilizado para explicar o recente fenômeno na web em que pessoas comuns descrevem conteúdos (notícias, blogs, vídeos, imagens, sites, etc) através de *tags* ou *labels* (palavras-chaves, não necessariamente presentes no conteúdo sendo descrito). Sites como Delicious (http://del.icio.us), Flickr (www.flickr.com), Youtube (www.youtube.com), WikiMapia (www.wikimapia.org) e outros permitem que usuários possam utilizar *tags* para descrever conteúdos para si ou para que outras pessoas possam recuperar tais conteúdos com maior facilidade posteriormente.

Uma Folksonomia é uma espécie de classificação (taxonomia) feita por usuários comuns e não por especialistas (neste último caso, seria chamada de taxonomia ou ontologia). Este termo também se contrapõe a "folk taxonomy" [11] porque não gera uma classificação estável e culturalmente aceita (segundo a Wikipedia).

O termo Folksonomia foi inicialmente utilizado por Thomas Vander Wal em um fórum de discussão [13]. Alguns termos sinônimos estão sendo utilizados tais como "*tags*onomia" e "*tag*ging" (ou "collaborative/social *tag*ging" ou "*tag* generation" ou "*tag*/web annotation"). Outro termo que poderia ser utilizado é "thesaurus social", pois uma folksonomia utiliza termos para descrever classes; o termo "social thesaurus" foi utilizado em [13] mas não é comum entre artigos científicos.

A principal diferença técnica de uma folksonomia para uma taxonomia é que a primeira não estabelece uma relação hierárquica entre as classes (no caso, as *tags*), nem exige exclusividade entre as classes (um elemento pode pertencer a mais de uma classe) [6].

Algumas vantagens do uso de Folksonomia são:

- servem para recuperação (filtro) de documentos e descoberta de conhecimento;
- servem para compartilhar conhecimento e são uma forma de organização [6].
- geram menor carga intelectual para associar classes pré-existentes; às vezes, é difícil classificar um conteúdo pertencente a vários assuntos [15]; exemplo: um documento sobre informática médica deveria ser classificado em "informática" ou "medicina" ou "saúde" ?; talvez o correto fosse classificá-lo em "informática médica" (mas esta classe pode não existir);
- é uma alternativa para a web semântica [6]; também chamada de "lowercase semantic web" [3], é uma abordagem evolucionária para gradualmente acrescentar significados simples a conteúdos e quebrar barreiras para reuso de informações;
- permite definir conteúdos secundários, enquanto taxonomias falam do assunto central ou principal;
- a criação é descentralizada, gerando uma estrutura menos rígida e mais flexível, enquanto que taxonomias são limitadas à hierarquia [1];
- já que o significado da *tag* é dado pelo coletivo (só vale dentro do contexto), conduz a simplicidade e facilidade de uso [1].

Porém existem também desvan*tag*ens no uso de *tags*:

- há problemas nas relações de significado entre *tags* e seus referentes [6]; como estudado pela Semiótica [2];
- há problemas de polissemia, sinonímia e variações lingüísticas (conjugações verbais, plurais, gênero) [6] e problemas com erros ortográficos e multi-palavras ou expressões (ex: SanFranciscoCalifornia) [1];
- há problemas no uso de *tags* genéricas versus *tags* específicas, assim como algumas *tags* são para interesse particular enquanto seu principal objetivo deveria ser o coletivo; a freqüência da *tag* pode indicar se é ela é genérica ou coletiva ou pessoal [6];

Apesar das desvantagens, o uso de *tags* para descrever conteúdo na Web é muito comum e está ganhando espaço. Entretanto, há poucos estudos sobre o processo de geração ou seleção das *tags* e, em geral, este processo é feito pelos usuários sem muitos critérios.

O objetivo deste artigo é estudar este processo, procurando encontrar padrões, ou seja, tentando identificar critérios, mesmo que não intencionais, no modo como os usuários selecionam tags. O foco está em descobrir padrões para poder auxiliar os usuários em processos futuros, sugerindo tags ou mesmo gerando palavras-chave automaticamente.

O artigo também se propõe a estudar *tags* selecionadas em consenso por um grupo de pessoas (mesmo sem discussão prévia ou conhecimento mútuo) e compará-las com *tags* que são escolhidas ou utilizadas por somente uma pessoa. Esta comparação poderá indicar algum critério melhor para seleção futura de *tags*, considerando o que se convencionou chamar de "inteligência coletiva".

Este artigo está estruturado da seguinte forma: na Seção 2 nós apresentamos os trabalhos relacionados e anteriores a esse artigo. A Seção 3 demonstra a proposta do trabalho e faz uma apresentação geral dos dados utilizados, dos experimentos e dos padrões analisados, concluindo apresentando o resultado dos principais padrões identificados. Por fim na Seção 4 apresentamos os principais resultados relacionando-os com suas aplicações e trabalhos futuros pretendidos.

2 Trabalhos Relacionados

Os trabalhos [15] e [6] apresentam categorias de *tags*, tais como "baseados em conteúdos", "de contexto (local e tempo)", "atributos (ex: origem ou fonte)", "subjetivos ou pessoais", "organizacionais (ex: "a ler")". Entretanto, tais trabalhos não indicam quais os tipos mais freqüentes.

Já [8] analisa a distribuição de *tags* ao longo do tempo. A conclusão é que os sites mais populares seguem um padrão no uso de *tags*: há uma preferência por reuso e por *tags* comuns. O trabalho de Cattuto [2] também identificou que os usuários tendem a usar mais freqüentemente *tags* que foram adicionadas recentemente do que *tags* mais antigas.

Em [6] também analisa-se a distribuição de *tags*, chegando a algumas conclusões:

a) há uma forte relação entre o número de conteúdos marcados por um usuário e o número de *tags* que este usa;

b) a lista de *tags* utilizadas por um usuário cresce com o tempo; e

c) a popularidade de uma *tag* pode variar com o tempo (crescer, estagnar, ser redescoberta).

Contudo, a geração de *tags* para a Web Semântica é um processo trabalhoso, que consome muito tempo e exige conhecimento de especialistas humanos. Por isto, eles propõem a geração automática de *tags* personalizadas a partir de documentos de interesse do usuário. [3]

O seguinte trabalho [15] define alguns critérios para geração adequada de *tags*: abrangência de diferentes tópicos presentes, popularidade, menor

esforço, uniformidade, uso de sinônimos. Além disto, um bom conjunto de *tags* deve incluir ambos os tipos genéricos e específicos. Uma estatística interessante descoberta no citado trabalho é que 92% de objetos na Web possuem 5 ou menos *tags* associadas, sendo que 79% têm 3 ou menos. Outra descoberta é que a distribuição das *tags* segue uma função Zipf e seu "Princípio do Menor Esforço" [5]. Assim o reuso de *tags* torna-se um dos métodos manuais mais comuns, principalmente por imitação a pessoas que já usaram as *tags* e que possuem boa reputação.

Já [1] chegou à conclusão que *tags* definidas manualmente são melhores para classificar textos em categorias mais gerais, enquanto que as *tags* escolhidas automaticamente são mais específicas. Além disto, as *tags* extraídas automaticamente do conteúdo dos textos são melhores que as selecionadas manualmente em processos de agrupamento por similaridade.

Os autores Brooks e Montanez [1] sugerem também extrair *tags* automaticamente do conteúdo do texto (analisando palavras mais relevantes) para minimizar o esforço do usuário. Entretanto, os mesmos autores afiram que esta escolha deve ser parametrizada pelo chamado "social *tag*ging", ou seja, o uso de *tags* pela coletividade gera melhores resultados, pois ajuda a diminuir problemas com sinônimos e variações lingüísticas: "*O uso coletivo produz um significado mais preciso do que definições feitas por uma única pessoa*" (Brooks e Montanez, 2006, p.626).

3 Proposta do Trabalho e Padrões Descobertos

O objetivo do presente trabalho é analisar a forma como as pessoas selecionam *tags*. Foram feitos experimentos para tentar identificar os padrões escondidos por trás de tal seleção, principalmente no que tange ao uso de palavras que aparecem no texto (título, descrição, no início do texto, etc.). O objetivo não é dizer qual o melhor método de escolha, mas sim identificar padrões na seleção ou escolha de *tags*, para que possam ser reutilizados por métodos automáticos ou que possam sugerir palavras-chave para usuários.

Também se deseja verificar se as pessoas utilizam como *tags* palavras que não aparecem nos textos sendo descritos. [12] consegue demonstrar que o uso de termos relacionados, mas que não estão presentes nos textos (técnica da expansão semântica), ajuda a conectar melhor páginas Web e anúncios (uso de palavras do mesmo contexto mas que não estão presentes num texto).

Buscamos também comparar padrões de uso individual e coletivo das *tags*. Uma suposição é que há diferença, ou seja, *tags* usadas por várias pessoas (coletivamente) seguem um padrão diferente das *tags* individuais (usadas por somente uma pessoa). O intuito é tentar descobrir algum tipo de inteligência coletiva, ou seja, padrões que se repetem entre as pessoas e que possam indicar uma escolha mais acertada. Recentemente, tem-se falado muito em aproveitar a chamada Inteligência Coletiva (ou Sabedoria das Massas) para gerar melhores resultados em diversos campos [10] e [14]. Este tipo de inteligência acredita que ninguém sabe tudo, mas todos sabem alguma coisa e que, portanto, a união das sabedorias individuais ou de pequenos grupos poderia resultar em decisões melhores até mesmo que as tomadas por especialistas. Supondo que haja diferença entre o padrão de uso individual de *tags* e o padrão coletivo e, supondo que a coletividade toma melhores decisões, os padrões encontrados na coletividade poderiam ser usados por sistemas automáticos para descrever conteúdos na Web.

3.1 Funcionamento do Delicious

Del.icio.us, ou Delicious [4] é um sistema de colaboração via *tags* para sites. Seu criador Joshua Schachter chama-o de gerenciador social de favoritos.

É composto basicamente de três entidades: usuário, *tag* e websites. Após o cadastro o usuário já pode incluir seus favoritos no site, todo favorito contém um título, uma descrição (opcional) e um conjunto de *tags* associado (também opcional). Após a inclusão o favorito inserido será mostrado na página do usuário (http://del.icio.us/username). Nessa página, além dos favoritos é mostrado o conjunto de *tags* utilizadas em algum momento pelo usuário.

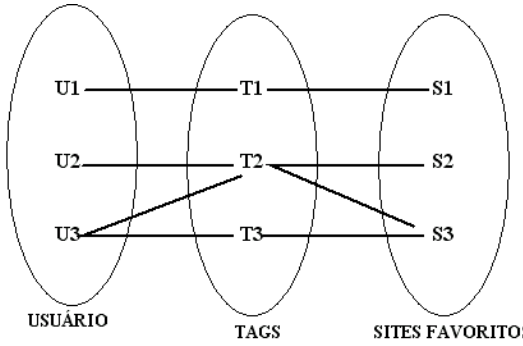

Figura 1: Gráfico de Representação das Entidades no Delicious

O site Delicious é considerado social porque o usuário pode ver não apenas os seus favoritos mas também navegar nas páginas dos demais usuários, na Figura 1 é possível observar essa relação. A grande vantagem é que com isso é possível formar um ranking das páginas populares (as que mais tiveram cadastros) e das *tags* mais populares, além de poder navegar entre elas. Por exemplo, a seguinte página no Delicious (http://del.icio.us/*tag*/folhaonline) irá encontrar todos os favoritos já incluídos por todos os usuários relacionados à *tag* "folhaonline". Foi dessa forma que conseguimos filtrar as notícias para as fontes que analisamos.

3.2 Dados Analisados

Nossas análises foram realizadas entre os dias 4 e 6 de maio de 2007 e englobaram um *corpus* de 797 notícias cadastradas no site Delicious (del.icio.us), consideradas

adequadas para análise e relacionadas a Folha de São Paulo[1] (568) e Estadão[2] (269). As notícias adequadas compreendem as que apresentaram pelo menos uma *tag* associada ao cadastro, das 1000 notícias verificadas 797 continham, portanto, essa restrição.

O objetivo do experimento é tentar descobrir padrões na escolha de *tags*, mesmo que estes padrões não sejam intencionais, isto é, que possam ser usados pelos usuários de forma inconsciente. Cabe salientar que as *tags*, neste caso, são *tokens* únicos, mesmo quando combinando expressões (ex: SanFranciscoCalifornia).

O experimento também se propõe a estudar *tags* selecionadas em consenso por um grupo de pessoas (mesmo sem discussão prévia ou conhecimento mútuo) e compará-las com *tags* que são escolhidas ou utilizadas por somente uma pessoa. Esta comparação poderá indicar algum critério melhor para seleção futura de *tags*, considerando o que se convencionou chamar de "inteligência coletiva".

O processo de análise abrange a recuperação das *tags* cadastradas para cada *URL* e a posterior verificação delas no conteúdo da notícia, levando em conta diferentes tipos de análise e formas de conteúdo, recuperados e tratados para tornar possíveis as comparações (*tags* e notícias).

O tratamento dos dados verificados foi necessário tanto para as *tags* como para o conteúdo das notícias. Para as *tags* foi necessária a retirada de caracteres especiais, por exemplo, ($,"¨´\"&()%¿), a retirada de acentos, a desconsideração de *tags* com separadores e a união de *tags* similares. Os separadores no contexto das *tags* representam geralmente palavras compostas, visto que no Delicious não é possível identificá-las com espaço. Uma análise de Golder et al. [7] mostrou que os principais caracteres separadores utilizados no Delicious são, respectivamente (-) hífen, (_) *Underline,*(/) Barra, (.) Ponto e (:) 2 Pontos. A presença desses caracteres nas *tags*, com exceção do hífen, representa palavras de improvável presença no texto das notícias. Devido a isso elas foram desconsideradas para a análise dos padrões. Após essas verificações todas as *tags* similares foram unidas e suas freqüências somadas; para tanto se utilizou o mesmo processo aplicado na verificação de freqüência das palavras o qual é melhor explicado mais abaixo.

Com relação às notícias, foi necessário normalizar seu conteúdo através da retirada de *stopwords* (palavras comuns em português e outros termos considerados irrelevantes para as comparações), retirada de *tags html* (com exceção de links) e de tratamentos específicos relacionados a cada tipo de análise.

Para a retirada das *tags html* os links tiveram tratamento especial pois se observou que tanto o endereço das *urls* como suas descrições poderiam conter potenciais *tags*. Dessa forma as *urls* foram separadas e unidas com sua descrição em cada posição do texto. O seguinte link: (Videos) por exemplo, seria normalizado da seguinte forma: (www youtube com videos), com a retirada das *stopwords* passaria para (youtube videos).

As análises desenvolvidas foram as seguintes:

Palavras Comuns. Representa uma análise simples, onde todas as palavras da notícia (normalizadas sem acento e com letra minúscula) são agrupadas em um *array* o qual mantém associado a respectiva posição delas no texto.

Substantivos. Similar a anterior com a diferença que através de análise sintática, todas as palavras da notícia são verificadas sendo extraídos somente os substantivos, sendo mantida a associação com a posição no texto representada pela palavra analisada.

Nomes Próprios. Nessa análise somente são consideradas as palavras em maiúsculo, da mesma forma que as anteriores, é mantida num *array* de termos que contém uma associação com a posição de cada palavra no texto. Um problema na verificação de Nomes Próprios pode ocorrer nas palavras iniciais de cada frase, estas embora em maiúsculo podem representar palavras irrelevantes para nossa análise (i. e. preposições, advérbios), com a retirada de *stopwords* conseguimos resolver o problema.

Os tipos de conteúdo verificados nas notícias em cada tipo de análise foram:

Título. As fontes escolhidas apresentavam *tags html* padronizadas para identificação do título, sendo possível sua captura. Diferente da descrição da notícia, a qual não foi possível verificar para ambos os sites (Folha e Estadão).

Frase. Como os caracteres separadores de frases são únicos (. ! ?) foi possível separá-las para a verificação.

Parágrafo. Cada site possuía *tags html* (

 ou <p>) padronizadas para todas as notícias responsáveis unicamente pra ser separar parágrafos, as quais foram utilizadas para captura.

Outro tipo de verificação foi a **Freqüência de Palavras**, a idéia era obter a posição (entre as mais freqüentes) que cada *tag* encontrada no texto tinha obtido. Para viabilizar essa análise foi necessário agrupar por similaridade as palavras do texto e contabilizar sua freqüência. Por questões de desempenho, implementamos uma verificação especial através dos seguintes passos:

1) Análise Especial das Palavras. Antes da Análise de Similaridade foi necessário verificar se as palavras comparadas poderiam ser consideradas similares. Assumiu-se que somente palavras com o último caracter diferente passariam para a etapa de verificação de

[1] http://www.folha.com.br

[2] http://www.estadao.com.br

similaridade. Esse procedimento foi adotado para resolver alguns problemas relacionados ao algoritmo de similaridade empregado. As palavras "filha" e "folha", por exemplo, sem esse tipo de verificação preliminar seriam consideradas similares segundo a verificação abaixo.

2) Verificação de Similaridade. Como as *tags* apresentam pouca variação léxica, principalmente relacionada ao plural, foi adotada uma verificação de similaridade especial. Para essa análise foi utilizada a distância Levenshtein, a qual é ideal para comparação de "strings" pequenas [9] apresentando menor complexidade que outros algoritmos estudados. Para ser considerada similar assumiu-se que duas "string*s*" deveriam ter grau de similaridade maior ou igual a 80%. Esse grau foi adotado porque evita que "strings" muito curtas (que apresentam muitas variações e similaridades errôneas) sejam consideradas similares. Dessa forma somente "string*s*" com mais que 4 caracteres podem ser consideradas similares. Por exemplo, a comparação de "casas" e "casa" retornaria similaridade de 80% visto que a distância de ambas é de 1 (4/5=0,8), o que não ocorre na comparação inversa ("casa" com "casas") embora a distância também seja 1 a primeira "string" tem apenas 4 caracteres e a similaridade seria (3/4=0,75). Como padrão é interessante notar que para as "string*s*" consideradas similares priorizou-se o armazenamento sempre da "string" de maior tamanho.

3.3 Padrões Descobertos

Análise Geral. Através da comparação entre as *tags* e o conteúdo da notícia podemos obter vários dados de âmbito geral, dentre eles convém citar: a quantidade média de palavras nas notícias (QMP = 286,68), a média de cadastros por notícia (MCN = 2,56), a porcentagem de notícias que tiveram alguma *tag* encontrada (PNTE = 74,27%), a média de *tags* cadastradas por notícia (MTCN = 5,14) e a porcentagem de *tags* encontradas (PTE=27,04%).

Para a PNTE completa-se a média de *tags* encontradas por notícia (MTEN = 1,63) o qual é a razão do número total de *tags* encontradas pelo número de notícias verificadas. E para a MTCN é importante citar a média de *tags* cadastradas para notícias que tiveram alguma *tag* encontrada (MTCE = 6,02) e a média de *tags* para notícias que não tiveram nenhuma *tag* encontrada no texto (MTCN = 2,59), o qual ilustra que o número de *tags* cadastradas é fator importante para sua existência na notícia.

Vale notar que para o cálculo da PNTE e da PTE também levaram-se em conta as *tags* similares encontradas no texto. As notícias que não tiveram *tags* no texto mas possuíam *tags* similares representaram 5% do total.

Análise dos Padrões. Para realizar a análise, as *tags* utilizadas foram separadas em 2 grupos: um grupo contendo as *tags* utilizadas por mais de uma pessoa para descrever a mesma notícia e um grupo de *tags* usadas por somente uma pessoa para descrever a mesma notícia.

No primeiro grupo (*tags* coletivas), foram analisadas 608 *tags*. Destas, 240 *tags* (39,47%) correspondiam a termos que também apareciam na notícia (no título ou no texto da notícia). Este baixo percentual indica que os usuários tendem a usar termos que não estão no texto, provavelmente utilizando-os para descrever conteúdo mais genérico (ex: assunto geral) ou mais específico (ex: interesses ou formas de organização pessoais).

No segundo grupo (*tags* individuais), foram analisadas 3278 *tags*. Destas, 811 *tags* (24,74%) correspondiam a termos que também apareciam na notícia (no título ou no texto da notícia). Este percentual é menor ainda do encontrando nas *tags* coletivas, confirmando a expectativa de que termos utilizados por mais pessoas tendem a ser extraídos do texto mais freqüentemente do que os termos utilizados por somente uma pessoa.

As *tags* encontradas nas notícias foram analisadas para tentar identificar padrões na seleção dos termos. A tabela 1 apresenta a proporção em que cada padrão aparece neste grupo. Pode-se notar que o padrão mais freqüente é utilizar palavras do primeiro parágrafo (42,75% das *tags* seguiam este padrão). O segundo padrão mais freqüente são as palavras presentes no título da notícia (41,64% das *tags* seguiam este padrão). Outro padrão bastante freqüente foi o uso de termos entre os 10 mais freqüentes no texto (38,29% das *tags* eram termos que estavam entre os 10 que mais apareciam no texto). Termos presentes na primeira frase do texto também foram comuns; 37,17% das *tags* escolhidas pelos usuários estavam na primeira frase.

A tabela 2 apresenta a proporção em que cada padrão aparece nas *811 tags* individuais encontradas nas notícias. Os padrões encontrados são semelhantes ao que se descobriu nas *tags* coletivas com algumas mudanças na ordem de freqüência. O padrão mais freqüente é utilizar palavras do título (34,75%), seguido por palavras do primeiro parágrafo (37,3%), pelos termos da primeira frase (33,17%) e depois pelas palavras entre as 10 mais freqüentes (23,21%).

Tabela 1: Padrões no uso de *Tags* coletivas (incluídas por mais de uma pessoa na mesma notícia)

Padrão	Palavras comuns	Substantivos	Nomes próprios
No Título	41,64%	39,78%	24,91%
Entre a 1a mais freqüente	12,27%	12,27%	10,41%
Entre as 3 mais freqüentes	21,19%	20,82%	17,10%
Entre as 5 mais freqüentes	27,88%	27,88%	21,93%

Entre as 10 mais freqüentes	38,29%	38,29%	29,74%
Na 1ª Frase	37,17%	36,80%	21,56%
Na 2ª Frase	17,84%	16,73%	8,92%
Nas outras frases	33,46%	31,23%	23,42%
No 1º Parágrafo	42,75%	41,64%	26,02%
No 2º Parágrafo	29,37%	29,00%	18,22%
Nos outros parágrafos	20,82%	20,45%	17,10%

Comparando-se as proporções entre *tags* coletivas (1º grupo) e *tags* individuais (2º grupo), nota-se que a proporção de *tags* coletivas encontradas no título é 19% maior que a proporção das *tags* individuais encontradas no título. Da mesma forma, a proporção no quesito "presente entre os 10 termos mais freqüentes" é 64% maior nas coletivas que nas individuais e a proporção das *tags* individuais entre a 1ª palavra mais freqüente é o dobro da proporção nas *tags* coletivas. Uma possível explicação ou conclusão é que há um certo padrão na coletividade, uma espécie de inteligência coletiva, que guia as pessoas a tomarem decisões de forma semelhante.

Embora a grande maioria das *tags* usadas não está no texto, a proporção é menor entre as coletivas, as quais apresentam 62% a mais de *tags* encontradas do que as individuais. Uma das prováveis explicações talvez seja o fato de as *tags* individuais apresentarem mais termos relacionados a conotações pessoais (adjetivos,lembretes, etc), o que não é útil para nossa análise. Dentre as *tags* individuais, por exemplo, alguma das palavras mais comuns foram "cool", "funny" e "fun". Esse fato também ilustra a existência de palavras em inglês na descrição das *tags*, o que pode trazer incoerências na verificação. Uma possível solução seria a análise da *tag* traduzida no texto.

Tabela 2: Padrões no uso de *Tags* individuais (incluídas por somente uma pessoa na mesma notícia)

Padrão	Palavras comuns	Substantivos	Nomes próprios
No Título	34,75%	32,44%	18,47%
Entre a 1a mais freqüente	5,95%	5,83%	3,16%
Entre as 3 mais freqüentes	13,85%	13,61%	9,11%
Entre as 5 mais freqüentes	16,65%	16,16%	10,69%
Entre as 10 mais freqüentes	23,21%	22,48%	14,46%
Na 1ª Frase	33,17%	31,11%	18,23%
Na 2ª Frase	16,89%	15,67%	8,51%
Nas outras frases	39,13%	37,55%	21,02%
No 1º Parágrafo	37,30%	34,63%	21,75%
No 2º Parágrafo	23,33%	21,51%	11,79%
Nos outros parágrafos	28,68%	27,70%	16,77%

A partir dos experimentos também foi possível verificar relações nas posições de presenças das *tags* nos textos, sendo as mais importantes relacionadas ao título. Dentre todas as *tags* (sem diferenciar coletivas de individuais), por exemplo, a porcentagem de vezes que a palavra aparecia no título e também na 1a frase foi de 62,5%, já no 1º parágrafo a relação foi de 66%. Dentre as *tags* coletivas que apareciam no título, a porcentagem que também estavam na 1a frase foi de 76,3%, já nas *tags* individuais ficou em 64%. As *tags* coletivas também possuíram vantagem na relação do título com as 3 mais freqüentes no texto da notícia obtendo 41,9% de relação, nessas as *tags* individuais ficaram com 23,9%.

3.4 Conclusões Preliminares sobre os Padrões

Em média são utilizadas 5 *tags* por notícia, sendo que destas, de 1 a 2 são encontradas no texto. Portanto, um sistema automático poderia, por exemplo, sugerir *tags* baseadas na notícia, sendo que em média o usuário usaria de 20 a 40% do que é sugerido (para não sugerir um número muito elevado).

O número de *tags* cadastradas nas notícias influencia a presença delas no texto. Foi observado que, quanto maior o número de *tags* definidas para uma notícia, mais relacionadas ao texto elas se tornam, corroborando a tendência dos padrões desse artigo de maior presença de *tags* coletivas no texto. As *tags* que não estavam presentes no texto da notícia estariam relacionadas, portanto, a conotações pessoais, a descrições de utilidade somente ao usuário que a cadastrou.

Os resultados dos padrões mostraram uma tendência de as *tags* coletivas estarem presentes no texto. No entanto em algumas posições como o título, a proporção entre *tags* individuais e coletivas se manteve similar. Os padrões com maior proporção entre *tags* coletivas foram relacionados à freqüência, respectivamente os presentes "entre a 1a mais freqüente" e também "entre as 10 mais freqüentes", o que comprova a credibilidade desse tipo de análise para obtenção de palavras relevantes no texto.

Na análise de relação de presença, o título e a primeira frase, juntamente com o título e o primeiro parágrafo obtiveram o maior grau de relacionamento. Mais uma vez as *tags* coletivas apresentaram vantagens quanto às individuais. Para um sistema de sugestão automática de *tags* essas relações poderiam ser utilizadas para otimizar o processo (evitando procurar na 1ª frase ou parágrafo quando achar uma *tag* no título, por exemplo).

A análise de substantivos provou-se útil na identificação de termos relevantes nas notícias, obtendo

porcentagens de presença no texto muito similares às palavras comuns. Um dos problemas enfrentados nessa análise foi a dificuldade em identificar palavras em inglês, fato esse que contribuiu para a diferença nos resultados. Na comparação com os nomes próprios pode-se comprovar, também, que os substantivos são mais eficientes para representação de termos relevantes nos textos.

4 Conclusão

A partir do trabalho e da observação dos padrões, percebeu-se que a presença das *tags* cadastradas para alguma notícia no site Delicious tem relação com o conteúdo presente na página representada. A ocorrência de *tags* no texto variou conforme a posição e o tipo de análise, os quais mostraram uma prevalência na presença das palavras no título e no primeiro parágrafo/frase, ambos relacionados ao grupo de *tags* coletivas analisados.

Embora a porcentagem de *tags* presentes nas notícias seja baixa (27,04), a relação de notícias com *tags* presentes no texto (74,27) e o número médio de *tags* encontradas por notícia (1,63) apresentaram boa porcentagem e freqüência. Portanto, um sistema automático poderia, por exemplo, sugerir *tags* baseadas na notícia, sendo que em média o usuário usaria de 20 a 40% do que é sugerido (para não sugerir um número muito elevado). Os tipos de conteúdos que poderiam ser utilizados e que apresentaram maior ocorrência no texto foram respectivamente, o título, a 1ª frase, o 1º parágrafo e entre as 10 palavras mais freqüentes. A análise de palavras comuns e de substantivos apresentaram os melhores resultados, com pouca diferença entre elas.

No futuro pretende-se melhorar esse processo através da análise de sinônimos (gerar *tags* com o mesmo significado para ser verificado no texto), análise de *thesaurus* (para os problemas de homonímia) e análise de idioma (traduzir *tags* em inglês e verificar tradução no texto).

Com os resultados dessa análise e a identificação da posição das palavras que possam melhor identificar o texto, uma aplicação que surge é a análise de similaridade entre documentos. Através de algoritmos de agrupamento (*clustering)* poderíamos, por exemplo, refazer e melhorar o trabalho de Brooks e Montanez [1] separando diretamente as palavras principais do texto, visando assim obter os melhores *clusterings*.

Uma outra aplicação possível é utilizar os padrões identificados neste artigo para gerar uma lista de *tags* por período, relacionando-as com os principais eventos ocorridos num determinado período de tempo, permitindo a criação de sistemas de análise de tendências para determinados tipos de conteúdo.

Agradecimentos

O presente trabalho foi realizado com o apoio do CNPq, uma entidade do Governo Brasileiro voltada ao desenvolvimento científico e tecnológico.

Referências

[1].BROOKS, C. H.; MONTANEZ, N. (2006) Improved annotation of the blogosphere via auto*tag*ging and hierarchical clustering. In: *International World Wide Web Conference – WWW*, May 2006, Edinburgh, Scotland, p.625-631.

[2].CATTUTO, C. (2006) Semiotic dynamics in online social communities. European Physical Journal, v.46, n.2, p.33-37.

[3].CHIRITA, P. A.; COSTACHE, S.; HANDSCHUH, S.; NEJDL, W. (2007) P-*TAG*: large scale automatic generation of personalized annotation *Tags* for the web. In *World Wide Web Conference*, Banff, Canada, May 2007.

[4].Del.icio.us. http://del.icio.us/

[5] ZIPF, G.K. (1949) Human Behaviour and the Principle of Least Effort. Addison-Wesley, Cambridge, Massachusets.

[6].GOLDER, S. A.; HUBERMAN, B. A. (2006) Usage patterns of collaborative *tag*ging systems. Journal of Information Science, v.32, n.2, p.198-208.

[7].GUY, M., TONKIN, E. (2006) Tidying up *tags*? D-Lib Magazine 12

[8].HALPIN, H.; ROBU, V.; SHEPHERD, H. (2007) The complex dynamics of collaborative *tag*ging. In *World Wide Web Conference,* Banff, Canada, May 2007. *http://www.dlib.org/dlib/january06/guy/01guy.html*

[9].LEVENSHTEIN, V. I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, p. 707–710.

[10].LÉVY, P. (1998) A inteligência coletiva: por uma antropologia do ciberespaço. São Paulo: ed. Loyola.

[11].MATHES, A. (2004) Folksonomies - cooperative classification and communication through shared metadata. Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, December 2004.

[12].RIBEIRO-NETO, B.; CRISTO, M.; DE MOURA, E. S.; GOLGHER, P. B. (2005) Impedance coupling in content-target advertising. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Bahia, Brazil, July 2005, p.496-500.

[13].SMITH, G. (2004) "Folksonomy: social classification." August, 2004. *http://atomiq.org/archives/2004/08/folksonomy_social_classification.html*

[14].SUROWIECKI, J. (2004). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Little, Brown.

[15].XU, Z.; FU, Y.; MAO, J.; SU. D. (2006) Towards the semantic web: collaborative *tag* suggestions. In *Collaborative Web Tagging Workhop, WWW Conference*, Edinburg, Scotland, May 2006.