Tags Coletivas: Analisando Padrões de Uso para o Suporte a Sistemas de Folksonomia

[1]Cleber Gouvêa, [1,2]Stanley Loh, [2]Luís Fernando Fortes Garcia

[1]Centro Politécnico,
Universidade Católica de Pelotas, Pelotas
RS Brasil
cleber@sindiq.com.br

[2]Universidade Luterana do Brasil, Porto Alegre,RS
Brasil
sloh@terra.com.br,
luis@garcia.pro.br

RESUMO

Com a popularização dos serviços voltados a anotação descentralizada de informações por meio de *tags* torna-se necessário o desenvolvimento de serviços que auxiliem a correta identificação de *tags* e consequentemente a relevância das anotações realizadas. O presente trabalho busca a partir da identificação de padrões relacionados às *tags* coletivas (incluídas por mais de uma pessoa para descrever determinado objeto no *site* Delicious) sugerir métodos para a sugestão automática de *tags*. Para avaliação buscamos verificar a qualidade dos clusters criados a partir das *tags* identificadas automaticamente (por meio das métricas de coesão e acoplamento) demonstrando assim a viabilidade da estratégia utilizada.

ABSTRACT

The growing use of annotations for web content demands systems that support the appropriated selection of *tags*. This work analyzes patterns in the use of collective *tags* (those registered by many people) with the goal of defining methods for the automatic selection or suggestion of *tags*. Each pattern generates a different automatic method. The methods are evaluated by the quality of clusters generated by the separation of contents according to *tags* selected by each method. The quality is evaluated through metrics as cohesion and coupling.

Palavras-Chaves

Folksonomia, *Tags*, Web Semântica, Inteligência Coletiva.

ACM Classificação

D.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, retrieval models.* Content Analysis and Indexing – abstracting methods, indexing methods. Online Information Services – Web-based services.

INTRODUÇÃO

Com a popularização na web dos serviços focados na colaboração (*user-created content*) variados sistemas de folksonomia têm surgido e proporcionado a descrição ou anotação descentralizada das informações por meio de *tags* (palavras-chave, não necessariamente presentes no conteúdo).

Sites como Delicious (http://del.icio.us), Flickr (www.flickr.com), Youtube (www.youtube.com) e WikiMapia (www.wikimapia.org) permitem aos usuários utilizarem *tags* para descrever conteúdos da Web, tais como notícias, páginas, imagens, vídeos e blogs. Estas *tags* funcionam como descritores ou indexadores para a classificação do conteúdo visando sua posterior recuperação, seja com intenção pessoal ou coletiva.

Como a inclusão das *tags* é realizada de forma manual, despendendo tempo para sua inclusão torna-se particularmente útil a identificação de métodos que auxiliem esse processo a partir da sugestão de *tags* relacionadas aos documentos. Esse fato é particularmente útil para determinados tipos de informações como as jornalísticas (onde embora alguns *sites* descrevam as informações por meio de categorias/assuntos ou mesmo *tags* esse processo despende tempo dos editores), outros *sites* embora sugiram *tags*, as recomendam levando em conta apenas as *tags* já associadas por usuários para a informação, não levando em conta o conteúdo do documento.

Embora alguns trabalhos já busquem identificar técnicas para a sugestão automática de *tags* [1][7], estes são no entanto focados em *blogs* e não apresentam uma análise particular relacionada aos textos jornalísticos.

Este trabalho busca, portanto auxiliar na resolução desses problemas levando em conta para isto os padrões identificados em um trabalho anterior [3] (onde verificou-se que *tags* coletivas, ou seja, as *tags* selecionadas em consenso por um grupo de pessoas no *site* Delicious, mesmo sem discussão prévia ou conhecimento mútuo, possuem mais freqüência no texto dos documentos por elas identificados que *tags* individuais, auxiliando assim na identificação de padrões de ocorrência no texto e na

sugestão de métodos para a sugestão automática de *tags*). Dessa forma buscamos testar estes padrões identificados e verificar assim sua viabilidade para a sugestão de *tags*.

Para ser eficiente a descrição da notícia deve resultar em *tags* que não somente façam a **descrição do conteúdo** de forma satisfatória, mas também realizem o **agrupamento de notícias relacionadas**. Por isto, para avaliação dos resultados, foi realizado o agrupamento das notícias de acordo com as *tags* selecionadas por cada método, sendo então verificada a qualidade dos agrupamentos por meio das métricas de coesão e acoplamento.

Este trabalho está estruturado da seguinte forma: primeiramente descrevemos em mais detalhes o problema abordado pelo trabalho, ilustrando em seguida os experimentos utilizados para sua resolução, complementando com os resultados alcançados e concluindo com um resumo das principais vantagens do trabalho e dos trabalhos futuros pretendidos.

DEFINIÇÃO DO PROBLEMA

Este trabalho procura principalmente resolver os problemas provenientes da inclusão manual de *tags*. Busca-se com isso auxiliar a sugestão de *tags* para sistemas de folksonomia os quais tem a carência de não levar em conta o conteúdo dos documentos na indicação de *tags* e ajudar também *sites* na web como notícias que embora utilizem o conceito de *tags* realizam o seu cadastramento de forma manual pelos editores, despendendo tempo e necessitando assim métodos automáticos para a sua inclusão. Para isso nos motivamos nas seguintes constatações identificadas por trabalhos anteriores.

Xu et al. [8] definiram critérios para a geração de *tags* adequadas: cobertura de diferentes tópicos, popularidade, menor esforço, uniformidade, uso de sinônimos. Verificando também que o reuso de *tags* é um dos métodos manuais mais utilizados para sua seleção, especialmente o uso de *tags* já utilizadas por pessoas com boa reputação. Já Cattuto [2] notou que os usuários tendem a usar com maior freqüência as *tags* adicionadas mais recentemente.

Brooks and Montanez [1] concluíram que as *tags* definidas manualmente são melhores para classificar textos em categorias enquanto que as *tags* automáticas são melhores para determinar categorias mais específicas (as manuais são melhores para categorias mais genéricas).

Uma investigação preliminar [3] concluiu que de forma geral 27% das *tags* associadas a notícias publicadas no *site* Delicious estavam presentes no texto das notícias, e que 74,2% das notícias tinham pelo menos uma das *tags* presente no texto, sendo que a proporção é maior entre as *tags* coletivas, as quais apresentam 62% a mais de presença que as *tags* individuais. Considerando que a média de *tags* cadastradas por notícia é de 5,14 *tags*, pode-se dizer que boa parte da tarefa de selecionar *tags* poderia se feita de forma automática.

As *tags* coletivas apresentaram de forma geral maior ocorrência no texto das notícias nos vários padrões analisados em [3] com destaque para os substantivos presentes no título, 1ª frase e TOP 3 palavras mais freqüentes da notícia. Na análise de relação de presença, o título e a primeira frase e o título e as TOP3 palavras mais freqüentes também apresentaram mais ocorrência para as *tags* coletivas que para as individuais.

O presente artigo busca considerar estes resultados para a sugestão de métodos de identificação de *tags* levando em conta a similaridade das notícias agrupadas por elas. O intuito é tentar descobrir se os padrões que envolvem algum tipo de inteligência coletiva (no caso relacionados às *tags* coletivas), ou seja, padrões que se repetem entre as pessoas podem indicar uma escolha mais acertada podendo ser usados por sistemas automáticos para descrever conteúdos na Web.

Com base nestas premissas, o presente trabalho selecionou os principais métodos apontados pelas *tags* coletivas visando testá-lo e compará-los, conforme descreve a próxima subseção.

Este trabalho propõe-se a estender a investigação de Brooks e Montanez [1] utilizando diferentes métodos de seleção automática de *tags*. Outro diferencial em relação ao trabalho de Brooks e Montanez é que estes avaliam a qualidade dos agrupamentos (*clusters*) gerados pelas *tags* apenas com a métrica de coesão, ou seja, verificando a similaridade entre os elementos do mesmo agrupamento. O presente trabalho utiliza, além da métrica de coesão, a métrica de acoplamento, visando avaliar também a similaridade entre os agrupamentos obtendo com isso resultados mais detalhados.

Métodos Investigados

Os seguintes métodos para a seleção de *tags* foram investigados:

- **Título:** esta técnica seleciona como *tags* os substantivos presentes no título da notícia;
- **1ª Frase:** esta técnica seleciona os substantivos presentes na 1a frase do texto da notícia;
- **Top 3:** esta técnica seleciona os 3 substantivos mais freqüentes no texto da notícia;
- **Título e 1ª Frase:** esta técnica seleciona os substantivos presentes tanto no título da notícia quanto na 1a frase do texto;
- **Título e TOP 3**: esta técnica seleciona os substantivos presentes tanto no título da notícia quanto entre os 3 substantivos mais freqüentes no texto;
- **1ª Frase e TOP 3**: esta técnica seleciona os substantivos presentes tanto na 1ª frase do texto da notícia quanto entre os 3 substantivos mais freqüentes neste texto;

Para a verificação do conteúdo de cada notícia, priorizou-se a extração somente dos substantivos, sendo esses recuperados por meio de software especial de análise sintática. Essa decisão foi tomada devido aos resultados obtidos em um experimento anterior [3].

Com relação aos textos das notícias foi necessário normalizar seu conteúdo a partir da retirada de *stopwords* (palavras comuns em português e outros termos considerados irrelevantes para a existência de *tags*), retirada de *tags html* (mantendo a descrição dos *links*) e de tratamentos específicos relacionados a cada tipo de análise.

EXPERIMENTOS

Para teste dos métodos, foi utilizado um *corpus* de 1000 notícias coletadas entre os dias 17 a 19 de julho de 2007, sendo 500 notícias do *site* Estadão (www.estadao.com.br) e 500 do site da Folha (www.folha.com.br). Os métodos de seleção de *tags* foram aplicados sobre cada uma das 1000 notícias do corpus de teste.

Para avaliar a qualidade das *tags* selecionadas por cada um dos métodos, foi realizado um processo de agrupamento das notícias por *tag*. Isto é, grupos (*clusters*) foram formados para cada *tag*. Nestes grupos, foram alocadas as notícias que continham a *tag* correspondente. Cada notícia poderia participar em mais de um grupo.

Após, foi verificada a qualidade dos grupos gerados utilizando duas métricas: coesão e acoplamento. A coesão mede a similaridade média entre os elementos de um mesmo grupo (faz o cálculo de similaridade entre cada par de notícia dentro do mesmo grupo e então calcula-se a média). Já o acoplamento mede a similaridade entre os grupos (faz o cálculo de similaridade entre cada par de grupo e então calcula-se a média geral dos grupos) [4]. Para cada método, foi calculado um valor de coesão e um valor de acoplamento.

$$\text{Coesão (P)} = \frac{\sum_{i>j} Sim(p_i, p_j)}{m(m-1)/2} \qquad \textbf{(1)}$$

Onde ***m*** é o número de notícias no cluster P e cada *p* um membro do cluster P.

$$\text{Acoplamento (P)} = \frac{\sum_{i>j} Sim(c_i, c_j)}{m(m-1)/2} \qquad \textbf{(2)}$$

Sendo *c* o centróide de determinado cluster presente em P e *m* o número de clusters presentes em P.

Como medida de similaridade adotou-se a função do cosseno [6].

A qualidade do agrupamento gerado é melhor quando há maior similaridade entre os elementos dentro de um grupo (maior coesão) e menor similaridade com os elementos de grupos diferentes (menor acoplamento). Por esta razão, as medidas de coesão e acoplamento foram combinadas numa só.

Após a recuperação das *tags* de cada tipo de conteúdo, os clusters foram construídos, sendo um por *tag*, onde cada cluster contém um vetor relacionando cada notícia que possui essa *tag*. Para a construção dos clusters com as associações foi necessário um processamento posterior a esse, dessa forma foi verificado as *tags* formadas em cada associação, sendo agrupadas as notícias compartilhadas pelos clusters anteriormente formados.

RESULTADOS

Os resultados dos experimentos relacionados a coesão e acoplamento médios para cada padrão são apresentados na tabela 1.

Tabela 1. Coesão, Acoplamento, Quociente e diferença para o melhor caso.

Análises	Coes.	Acopl.	Coes. / Acopl.	Dif. (%)
Título ∩ TOP 3	0,395	0,080	4,937	-
1ª Frase ∩ TOP 3	0,399	0,081	4,925	**-0,2**
Top 3	0,367	0,086	4,267	**-15,7**
Título ∩ 1ª Frase	0,352	0,083	4,240	**-16,4**
Título	0,303	0,087	3,482	**- 41,7**
1ª Frase	0,268	0,096	2,791	**-76,8**

Como a coesão e o acoplamento de forma isolada não apresentam resultados confiáveis (1 cluster pode possuir coesão e acoplamentos altos) calculamos também o quociente da coesão pelo acoplamento. Como ilustrado esse cálculo demonstrou-se útil na comparação de alguns conteúdos, como é o caso da análise da intersecção da 1ª Frase com o TOP 3, o qual embora apresente maior coesão que a relação do Título com o TOP 3, acaba demonstrando resultados um pouco melhores quando relacionamos ambas as métricas.

No entanto, apenas o valor da coesão e acoplamento não são muito informativos, para ajudar a demonstrar a legitimidade dos resultados obtidos é necessário saber quais os limites inferiores dos clusters que o sistema poderia apresentar. Para isso realizamos uma análise onde notícias foram incluídas em clusters aleatórios sendo computada a média da similaridade do cosseno apresentada, onde obtemos coesão média de 0.047 e acoplamento de 0.232. Outra análise que pode ajudar a validar os resultados foi a realizada em [1], a qual obteve coesão média de 0.4 para os tópicos considerados similares pelo *site* Google News (news.google.com).

CONCLUSÃO E TRABALHOS FUTUROS

O trabalho buscou analisar a possibilidade de identificação de padrões no uso de *tags* coletivas e a possível utilização deles para a sugestão de métodos automáticos para a identificação de *tags*.

A partir da análise da qualidade dos clusters criados pelos métodos analisados (baseados nos padrões identificados por [3] para as *tags* coletivas) verificamos o agrupamento de notícias obtendo resultados próximos aos alcançados por [1] para a coesão dos tópicos similares apresentados no Google News, demonstrando assim a viabilidade dos métodos propostos.

Com base na constatação que as *tags* coletivas auxiliam na identificação de métodos para a sugestão automática de *tags*, uma importante aplicação é a utilização dessa idéia por sistemas de folksonomia de variados tipos, buscando a partir da identificação de padrões da posição das *tags* no textos sugerir novas *tags* baseadas no conteúdo dos documentos, não só de notícias mas de variados tipos de informação, aumentando assim a variedade de *tags* associadas.

Especificamente com relação à anotação de notícias o trabalho possibilitou a sugestão de métodos para a sua descrição dinâmica por meio da identificação automática de *tags*. Isto pode facilitar o trabalho de editores de *sites* de notícias que precisam descrever as informações manualmente, não precisando despender tempo analisando seu conteúdo para subjetivamente identificar os termos principais. As *tags* sugeridas poderiam também ser associadas com a data da publicação das notícias relacionadas a elas, possibilitando assim a navegação temporal e o relacionamento de eventos ocorridos ao longo do tempo.

Para a otimização dos resultados, especificamente com relação à sugestão automática de *tags* alguns dos trabalhos futuros pretendidos são os seguintes:

- Buscar agrupar *tags* analisando o radical das palavras (stemming) visando obter resultados mais precisos.
- Testar métodos para a verificação de *tags* a partir de um sistema de identificação de Nomes Próprios nos textos.
- Testar a possibilidade da identificação de *tags* a partir dos valores relacionados ao ponto de transição (transition point), o qual baseia-se no pressuposto que termos com freqüência média tendem a definir de forma mais precisa o conceito central dos documentos, sendo essa freqüência identificada por meio de uma fórmula específica definida em [5].

AGRADECIMENTOS

Este trabalho é parcialmente apoiado por CNPq, FAPERGS e CAPES.

Cleber Gouvêa é mestrando em Ciência da Computação pela Universidade Católica de Pelotas, no Brasil. Trabalha desde 2005 com projetos envolvendo folksonomia e recuperação de informações na web (*Tagging e Geotagging*).

Stanley Loh é professor da Universidade Católica de Pelotas e da Universidade Luterana do Brasil, no Brasil. Ele também trabalha na Intext Mining Ltda, uma companhia brasileira que desenvolve tecnologias para a análise de textos. Ele também é doutor em Ciência da Computação, obtido em 2001 na Universidade do Rio Grande do Sul. Ele tem realizado pesquisas em sistemas de recomendação, *data-text-web mining* e tecnologia aplicadas a gestão do conhecimento.

Luís Fernando Fortes Garcia é professor da Universidade Luterana do Brasil, Professor da Faculdade Dom Bosco de Porto Alegre no Brasil. Tem experiência na área de Ciência da Computação, com ênfase em Sistemas de Computação. Atuando principalmente nos seguintes temas: Hipermidia Adaptativa, Dispositivos Móveis, Internet Móvel, Sumarização. Possui doutorado em Ciências da Computação pela Universidade Federal do Rio Grande do Sul.

REFERÊNCIAS

[1] Brooks, C. H.; Montanez, N., Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: International World Wide Web Conference – WWW, May 2006, Edinburgh, Scotland, p.625-631.

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

[3] Gouvêa, C.; Loh, S.; Garcia, L. F. F., Folksonomias: Identificação de Padrões na Seleção de *Tags* para Descrever Conteúdos. In: Webmedia - Brazilian Symposium on Multimedia and the Web, 2007, Gramado, RS. Anais Brazilian Symposium on Multimedia and the Web, 2007. p. 9-12.

[4] Kramer, S.; Kaindl, H., Coupling and cohesion metrics for knowledge-based systems using frames and rules. ACM Trans. Softw. Eng. Methodol., New York, NY, USA, v.13, n.3,p.332–358,2004

[5] Pinto D., H. Jimenez-Salazar, P. Rosso, and E. Sanchis., Buap-upv tpirs: A system for document indexing reduction at webclef. In S. Verlag, editor, Accessing Multilingual Information Repositories, Revised Selected Papers CLEF05, volume 4022, pages873–879, 2006.

[6] Salton G. and McGill M. J., An Introduction to Modern Informartion Retrieval. Mcgraw-Hill, Inc., New York, 1983.

[7] Sood S., Hammond K., et al.. Birnbaum.TagAssist: Automatic *Tag* Suggestion for Blog Posts. In *International Conference on Weblogs and Social Media*, 2007.

[8] Xu, Z.; Fu, Y.; Mao, J.; Su. D., Towards the semantic web: collaborative *tag* suggestions. In Collaborative Web Tagging Workhop, WWW Conference, Edinburg, Scotland, May 2006.