Recomendação Associativa de Tags na Ausência de Informação Prévia

Eder Ferreira Martins
Jussara M. Almeida (orientadora), Marcos André Gonçalves (co-orientador)

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

{ederfm, jussara, mgoncalv}@dcc.ufmg.br

1. Motivação para a Pesquisa da Dissertação

Várias aplicações da Web 2.0 tornaram-se extremamente populares devido principalmente ao forte estímulo à criação e compartilhamento de conteúdo pelos próprios usuários e ao estabelecimento de comunidades online e redes sociais. Esse aumento de popularidade propiciou a criação de ricas coleções de dados nessas aplicações. Em sua maioria, esses dados são compostos por mídias ricas (e.g., vídeos e imagens), o que cria desafios para serviços de recuperação de informação (RI). Isto porque as técnicas de RI baseadas em conteúdo multimídia existentes apresentam problemas de escalabilidade ao grande tamanho das coleções existentes e às altas taxas de *upload*, bem como de robustez à baixa qualidade comumente associada ao conteúdo multimídia compartilhado nestas aplicações. Assim, muitos dos serviços de RI existentes exploram apenas o conteúdo textual, notadamente *tags*, comumente associado ao conteúdo multimídia.

Cada instância de conteúdo em uma aplicação da Web 2.0 é composta por um objeto principal, que pode estar armazenado em diferentes tipos de mídia (texto, áudio, vídeo, imagem) e por diversas fontes de informação associadas ao objeto, denominadas *atributos* do objeto. Os atributos de um objeto podem ser de diferentes tipos (e.g., textuais, sociais). Neste trabalho focamos em atributos textuais, que são blocos de texto frequentemente associados pelos usuários aos objetos.

Na maioria das aplicações Web 2.0, *tags* se destacam dentre todos os atributos textuais visto que elas geralmente descrevem razoavelmente bem o conteúdo dos objetos a que são associadas [Figueiredo et al. 2013, Li et al. 2008], assim como os interesses dos usuários que as utilizam, provendo assim uma melhor organização e descrição do conteúdo. De fato, estudos recentes mostram que *tags* são um dos melhores atributos para dar suporte a serviços de classificação automática [Figueiredo et al. 2013], busca [Li et al. 2008], e recomendação de conteúdo [Guy et al. 2010].

Nesse contexto, serviços de recomendação de *tags* ajudam a melhorar a qualidade tanto das *tags* disponíveis quanto dos sistemas de RI que as exploram como fontes de informação. A literatura contém uma grande variedade de estratégias de recomendação de *tags*, sendo que as que exploram padrões de coocorrência com *tags* previamente associadas aos objetos do sistema (incluindo *tags* disponíveis no próprio objeto alvo da recomendação), chamados de métodos associativos, têm produzido consistentemente resultados considerados estado da arte. Entretanto, a maioria dos métodos associativos de recomendação de *tags* assume a existência de um conjunto inicial de *tags* [Heymann et al. 2008, Sigurbjörnsson and van Zwol 2008, Menezes et al. 2010, Belém et al. 2011], não tendo sido avaliados em cenários onde o objeto alvo da

recomendação não tem nenhuma *tag*. Nesse cenário, a eficácia desses métodos pode ser muito prejudicada já que não podem utilizar os padrões de coocorrência.

Tal cenário, que aparece em uma parcela não desprezível (cerca de 18%) dos objetos em aplicações populares da Web 2.0 [Figueiredo et al. 2013], é uma variação de um problema conhecido em sistemas de recomendação chamado *cold start* [Schein et al. 2002]. Ele é o foco desta dissertação. Especificamente, abordamos o problema de recomendar *tags* para objetos que não possuem *tags* iniciais, refletindo um cenário em que o usuário está adicionando um novo objeto ao sistema, já preencheu alguns dos atributos textuais e precisa de sugestões de termos relevantes para usar como *tags*. O nosso objetivo é avaliar a robustez de métodos associativos estado da arte (que alcançam notória eficácia em cenários sem *cold start*) [Belém et al. 2011] em cenários de *cold start* e propor soluções que tornem tais métodos mais robustos neste cenário.

2. Resumo das Principais Contribuições

De modo resumido, as principais contribuições da dissertação são:

- Avaliação de vários métodos associativos estado da arte de recomendação de *tags* em um cenário de *cold start*, que ocorre em parcelas significativa dos objetos na Web 2.0.
- Proposição de estratégias de filtragem automática para gerar um conjunto alternativo de *tags* de modo a amenizar o *cold start* sobre uma família de métodos associativos.
- Proposição de métodos baseados em retroalimentação de relevância (*relevance feedback*) e programação genética que produzem recomendações de qualidade superior a métodos encontrados na literatura.
- Proposição de um arcabouço de avaliação da robustez de métodos baseados em *relevance feedback* à falta de cooperação do usuário.
- Extensa avaliação dos métodos propostos, comparando-os com várias alternativas, em diferentes cenários, incluindo cenários com baixa cooperação do usuário.

Os resultados diretos dessa dissertação produziram um artigo [Martins et al. 2013] na principal conferência nacional na área de Web, premiado como *melhor artigo*. Além disso, um artigo foi submetido a um dos principais periódicos da área, *Journal of the Association for Information Science and Technology* (JASIST, Qualis A1), que encontra-se atualmente aguardando publicação (*accepted for publication*). Como resultados indiretos, que contribuíram para a dissertação, vale ressaltar a participação do mestre em três artigos ([Belém et al. 2011, Belém et al. 2010, Belém et al. 2013]) nas principais conferências internacionais de Recuperação de Informação, duas *Qualis* A1 e uma A2, com uma taxa de aceitação média de 17%, e em um artigo em periódico internacional Qualis A2 ([Belém et al. 2014]).

3. Detalhamento dos Métodos Propostos

O estudo realizado pode ser subdividido em várias etapas, cada uma contendo uma contribuição original da nossa pesquisa. A primeira etapa consistiu na quantificação do impacto do *cold start* sobre uma família de métodos estado da arte baseados em padrões de coocorrência de *tags* considerando quatro grandes bases de dados coletadas de aplicações populares da Web 2.0 (Bibsonomy, Last.FM, YahooVideo e YouTube). Nosso estudo revelou que a eficácia desses métodos é reduzida em até 84% em tal cenário. Essa

perda é tão grande que, nesse cenário, os métodos de recomendação associativos são superados inclusive por métodos mais simples que não exploram padrões de coocorrência, diferentemente do que ocorre quando há *tags* previamente associadas ao objeto alvo da recomendação [Belém et al. 2011]. Tal quantificação é inédita na literatura e motiva o desenvolvimento de técnicas para amenizar o impacto do *cold start* sobre estes métodos.

Na segunda parte do trabalho, foram exploradas estratégias de filtragem de modo a construir um conjunto alternativo de *tags*, as quais seriam usadas como entrada para os métodos associativos para obter padrões de coocorrência. Para tal, o trabalho focou nos métodos associativos estado da arte propostos em [Belém et al. 2011]. Foram avaliadas diversas estratégias para gerar o conjunto inicial de *tags* pela filtragem tanto de termos de objetos similares no conjunto de treino quanto de termos obtidos de outros atributos textuais do objeto alvo da recomendação (e.g., título e descrição). Entretanto os ganhos dessas estratégias são limitados, o que nos motivou a buscar novas soluções para o problema.

A principal contribuição da dissertação é uma solução para reduzir o impacto do *cold start* em métodos associativos que explora as preferências por *tags* específicas manifestadas pelo usuário durante o processo de recomendação. Tais preferências são tratadas como um retorno implícito sobre a relevância das *tags* (*relevance feedback*). A ideia geral da estratégia proposta pode ser descrita em 3 passos: (i) um conjunto inicial de *tags* é apresentado para o usuário; (ii) o usuário seleciona *tags* relevantes nesse conjunto; (iii) as *tags* selecionadas como relevantes são usadas como entrada para o método de recomendação associativo enquanto que as *tags* não relevantes (não selecionadas) são inseridas em uma lista negra para prevenir que elas sejam recomendadas novamente nas próximas iterações. Tal processo é repetido até que o usuário decida parar de adicionar *tags* ao objeto (ou até um número máximo de iterações). Observe que a estratégia proposta leva em conta tanto as *tags* que foram selecionadas pelo usuário (*feedback* positivo), quanto as que não foram (*feedback* negativo) e requer apenas um pequeno esforço extra por parte do usuário[1] que é recompensado por uma grande melhoria na qualidade das *tags* recomendadas.

Foram propostas algumas variações da estratégia mencionada acima, criadas a partir da aplicação de diferentes métodos nos passos (i) e (iii). Em particular, foram propostos o uso de simples heurísticas bem como a aplicação de Programação Genética (PG) [Poli 2002] para gerar funções específicas para cada iteração. A adoção de PG ao problema se deve à sua capacidade de descobrir soluções quase-ótimas em grandes espaços de busca (que é o caso aqui) e grande sucesso no tratamento de diferentes problemas de RI, incluindo recomendação de *tags* [Belém et al. 2011]. PG é um processo iterativo que implementa um mecanismo de busca global. A ideia básica da PG é gerar uma solução mais eficiente ou eficaz para um problema, a partir de um conjunto inicial de possíveis soluções, que são combinadas utilizando operações inspiradas no processo evolucionário biológico. No arcabouço de PG, cada possível solução, chamada indivíduo, é representada por uma árvore composta por métricas relacionadas ao problema. Em cada geração do processo evolucionário, os indivíduos são avaliados de acordo com uma função de aptidão, também relacionada ao problema. O processo é repetido até que um valor de aptidão dado como meta seja atingido ou um número máximo de gerações seja alcançado.

Os resultados experimentais obtidos mostraram que a melhor estratégia proposta,

[1]Tipicamente, isso envolve apenas clicar em 2 ou 3 *tags* recomendadas, que é um esforço esperado para um usuário de um sistema de recomendação de *tags*.

que combina o uso de *relevance feedback* (positivo e negativo) com programação genética (chamada na dissertação de $PG + RF$) reduz efetivamente o impacto do *cold start*, superando em até 58% outros métodos estado da arte tomados como referência. Mais ainda, verificou-se, também, que a estratégia proposta supera em até 43% os métodos baseados em coocorrência mesmo em cenários nos quais o problema de *cold start* não ocorre.

Como última contribuição, foi também investigada a robustez da estratégia proposta à falta de cooperação do usuário. Isto é, foi avaliado o impacto na eficácia das recomendações da adição de ruído (e.g., termos irrelevantes) e do não assinalamento de termos relevantes por parte do usuário. Os resultados mostraram que a solução proposta permanece tão boa quanto, se não melhor que, os métodos de referência mesmo quando 10% do *feedback* provido pelo usuário seja composto por ruído ou se o usuário seleciona apenas 50% das *tags* relevantes apresentadas durante o processo de *feedback*.

4. Contextualização em Relação aos Trabalhos Relacionados

Cold start, i.e., a ausência de informação sobre novos usuários ou itens que evita que eles sejam recomendados, é um problema bem conhecido em sistemas de recomendação de itens [Preisach et al. 2010, Ness et al. 2009, Schein et al. 2002, Said et al. 2009, Givon and Lavrenko 2009, Bobadilla et al. 2012, Sun et al. 2011]. Muitas das abordagens existentes para lidar com tal problema se baseiam em algoritmos de aprendizado de máquina para melhorar filtros colaborativos [Bobadilla et al. 2012]. Alguns trabalhos exploram atributos textuais (e.g., *tags*) associados ao objeto alvo da recomendação como uma fonte alternativa de informação [Givon and Lavrenko 2009, Said et al. 2009], enquanto outros combinam o uso de atributos textuais e filtros colaborativos para minimizar o problema do *cold start* [Sun et al. 2011].

Em contraste, existem poucos trabalhos que tratam do problema do *cold start* no contexto específico de recomendação de *tags*. Ness *et al.* [Ness et al. 2009] descrevem uma técnica baseada na análise do conteúdo de áudio que ajuda a melhorar a eficácia de um sistema de recomendação de *tags* para conteúdo musical frente ao *cold start*. Este tipo de método, entretanto, requer algoritmos especializados para cada tipo de mídia (e.g., imagens, vídeos), o que tipicamente tem uma alta complexidade computacional. Preisach *et al.* [Preisach et al. 2010] propõem um algoritmo semi supervisionado puramente baseado em grafos para realizar recomendações personalizadas. Esta abordagem visa personalização, que foge um pouco ao foco da dissertação, e não utiliza padrões de coocorrência. Logo, ela não pode ser diretamente comparada à nossa solução.

Muitos trabalhos tratam do problema de recomendação de *tags* assumindo a existência de um conjunto inicial de *tags* associadas ao objeto alvo. Dentre esses métodos, aqueles que exploram padrões de coocorrência, chamados métodos associativos, têm alcançado consistentemente resultados superiores às alternativas [Sigurbjörnsson and van Zwol 2008, Heymann et al. 2008, Menezes et al. 2010, Belém et al. 2011]. Por exemplo, em [Sigurbjörnsson and van Zwol 2008], os autores propõem a aplicação de métricas globais de coocorrência de termos (e.g., confiança) para produzir uma ordenação das *tags* por relevância. Belém *et al.* [Belém et al. 2011] estende o trabalho de [Sigurbjörnsson and van Zwol 2008] por meio da aplicação de métricas de relevância de *tags* a termos extraídos de múltiplos atributos textuais do objeto alvo da recomendação. Esses, assim como outros métodos anteriores [Heymann et al. 2008,

Menezes et al. 2010], foram avaliados apenas para objetos contendo algumas *tags* iniciais. Duas contribuições desta dissertação são a avaliação de tais métodos em cenários de *cold start* e a proposição de novos métodos mais robustos em tal cenário.

Em outra direção, alguns trabalhos não exploram *tags* previamente associadas ao objeto alvo [Lipczak et al. 2009, Graham and Caverlee 2008]. Por exemplo, o método CTTR [Lipczak et al. 2009] extrai termos de outros atributos textuais do objeto alvo da recomendação, expande esses termos e os ordena pelo seu uso como *tags* em um conjunto de treino. Já o método Plurality [Graham and Caverlee 2008] combina um modelo vetorial com *relevance feedback* (RF) provido por usuários. Em particular, ele é o único método anteriormente proposto que explora RF para recomendação de *tags*. Porém, diferentemente de nossa estratégia, o Plurality não explora o *feedback* negativo, focando apenas no *feedback* positivo. Tanto o CTTR quanto o Plurality não exploram as *tags* previamente associadas ao objeto, podendo assim ser mais robustos ao *cold start*. Entretanto, nossa avaliação experimental mostrou que nossa abordagem supera em até 83% tanto o CTTR quanto o Plurality em várias coleções de dados. Vale ressaltar que outros trabalhos demonstraram o valor de utilizar ambos os tipos de *feedback* em outros contextos de RI [Ferreira et al. 2011]. Assim, o uso conjunto dos *feedbacks* positivo e negativo para lidar com o *cold start* em sistemas de recomendação de *tags* é uma contribuição original de nosso trabalho. Mais ainda, a análise da robustez dos métodos de recomendação à falta de cooperação ou erros do usuário é inexistente na literatura.

5. Conclusões e Trabalhos Futuros

Os resultados de pesquisa descritos na dissertação têm um grande potencial de impacto devido à sua aplicabilidade na melhoria da qualidade da informação em aplicações da Web 2.0 e de seus serviços de RI. Nossas principais contribuições incluem: (1) extensa avaliação de vários métodos estado da arte de recomendação de *tags* em um cenário de *cold start* que ocorre em muitos objetos na Web 2.0; (2) proposição de métodos baseados em *relevance feedback* positivo e negativo e programação genética que produzem recomendações de qualidade superior a métodos encontrados na literatura em cenários com e sem *cold start*; (3) proposição de um arcabouço de avaliação da robustez de métodos baseados em *relevance feedback* à falta de cooperação ou erros do usuário. Trabalhos futuros incluem a proposição de estratégias mais flexíveis para lidar com o *feedback* negativo bem como com a personalização das recomendações.

Referências

Belém, F., Martins, E., Almeida, J., and Gonçalves, M. (2013). Exploiting novelty and diversity in tag recommendation. Proc. ECIR.

Belém, F., Martins, E., Almeida, J., and Gonçalves, M. (2014). Personalized and object-centered tag recommendation methods for web 2.0 applications. *IP&M*, 50:524–553.

Belém, F., Martins, E., Almeida, J., Gonçalves, M., and Pappa, G. (2010). Exploiting co-occurrence and information quality metrics to recommend tags in web 2.0 applications. Proc. CIKM.

Belém, F., Martins, E., Pontes, T., Almeida, J., and Gonçalves, M. (2011). Associative tag recommendation exploiting multiple textual features. In *Proc. ACM SIGIR*.

Bobadilla, J., Ortega, F., Hernando, A., and Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge Based Systems*, 26:225–238.

Ferreira, C., dos Santos, J., Torres, R., Gonçalves, M. A., Rezende, R., and Fan, W. (2011). Relevance feedback based on genetic programming for image retrieval. *Pattern Recognition Letters*, 32(1):27–37.

Figueiredo, F., Pinto, H., Belém, F., Almeida, J. M., Gonçalves, M. A., Fernandes, D., and de Moura, E. S. (2013). Assessing the quality of textual features in social media. *Information Processing & Management*, 49(1):222–247.

Givon, S. and Lavrenko, V. (2009). Predicting social-tags for cold start book recommendations. In *Proc. ACM RecSys*.

Graham, R. and Caverlee, J. (2008). Exploring feedback models in interactive tagging. In *Proc. Conference on Web Intelligence and Intelligent Agent Technology*.

Guy, I., Zwerdling, N., Ronen, I., Carmel, D., and Uziel, E. (2010). Social Media Recommendation Based on People and Tags. In *Proc. ACM SIGIR*.

Heymann, P., Ramage, D., and Garcia-Molina, H. (2008). Social Tag Prediction. In *Proc. SIGIR*.

Li, X., Guo, L., and Zhao, Y. E. (2008). Tag-based Social Interest Discovery. In *Proc. WWW*.

Lipczak, M., Hu, Y., Kollet, Y., and Milios, E. (2009). Tag Sources For Recommendation In Collaborative Tagging Systems. In *Proc. ECML PKDD*.

Martins, E., Belém, F., Almeida, J., and Gonçalves, M. (2013). Measuring and Addressing the Impact of Cold Start on Associative Tag Recommenders. In *Proc. WebMedia*.

Menezes, G., Almeida, J., Belém, F., Gonçalves, M., Lacerda, A., Moura, E., Pappa, G., Veloso, A., and Ziviani, N. (2010). Demand-Driven Tag Recommendation. In *Proc. ECML PKDD*.

Ness, S., Theocharis, A., Tzanetakis, G., and Martins, L. (2009). Improving Automatic Music Tag Annotation using Stacked Generalization of Probabilistic SVM Outputs. In *ACM Multimedia*.

Poli, R. (2002). *Foundations of genetic programming*. Springer-Verlag New York, Inc.

Preisach, C., Marinho, L. B., and Schmidt-Thieme, L. (2010). Semi-Supervised Tag Recommendation - Using Untagged Resources to Mitigate Cold-Start Problems. In *PAKDD (1)*, volume 6118, pages 348–357.

Said, A., Wetzker, R., Umbrath, W., and Hennig, L. (2009). A Hybrid PLSA Approach for Warmer Cold Start in Folksonomy Recommendation. In *Proc. ACM RecSys*.

Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. (2002). Methods and Metrics for Cold Start Recommendations. In *Proc. ACM SIGIR*.

Sigurbjörnsson, B. and van Zwol, R. (2008). Flickr Tag Recommendation Based on Collective Knowledge. In *Proc. WWW*.

Sun, D., Luo, Z., and Zhang, F. (2011). A Novel Approach for Collaborative Filtering to Alleviate the New Item Cold-Start Problem. In *Proc. ISCIT*.