Escola Politécnica da Universidade de São Paulo
PMR2500 – Projeto de Conclusão do Curso I

Relatório de Revisão Bibliográfica

Nome	*Número USP*
Antônio Guilherme Ferreira Viggiano	6846450
Fernando Fochi Silveira Araújo	5894546

Orientador
Prof. Dr. Fábio Gagliardi Cozman

19 de março de 2014

Sumário

1 Objetivo do trabalho

O objetivo do presente Trabalho de Conclusão de Curso é o desenvolvimento de um Sistema de Recomendação de produtos para lojas de comércio online, com foco na sugestão feita através de marketing via e-mail.

2 Desafios científico-tecnológicos Metodologia

Um dos maiores desafios tecnológicos dos sistemas de recomendação é, atualmente, o da escalabilidade [1]. O sistema de recomendação deverá ser flexível no sentido de poder operar igualmente bem tanto em conjuntos pequenas quanto em grandes bases de dados, que podem chegar até centenas de milhões de clientes [2] e de produtos [3]. Isso significa que as recomendações devem ser suficientemente rápidas para poderem operar em tempo real e ainda assim proverem sugestões valiosas aos usuários.

O sistema de recomendação também deve prever quando enviar uma determinada recomendação, e não agir apenas mediante requisição do cliente [4]. É interessante, por exemplo, enviar recomendações de produtos com descontos a usuários que estão há algum tempo inativos no site, para que eles retornem a comprar. Da mesma forma, um sistema inteligente poderia sugerir produtos infantis a um usuário detectado como recém-casado.

Outro desafio científico ainda em estágio inicial de pesquisa é referente à diversidade das recomendações realizadas, também chamado de excesso de especialização [5]. Ao mesmo tempo que o sistema deve apresentar itens similares ao que o usuário está procurando, ele também deve sugerir itens que o usuário desconheça ou que nem saiba que poderiam interessá-lo.

Por fim, um desafio científico que este trabalho enfrentará é a execução de um sistema híbrido do ponto de vista de efemeridade e persistência, ao construir um modelo de recomendação que integre entre as preferências de curto e longo termo dos usuários [6]. A análise dos dados de compras anteriores, bem como de dados demográficos, deverá portanto ser incorporada à análise de característica dos produtos, a fim de enriquecer a acurácia do sistema [1].

Esse tópico de pesquisa inclui ainda diversos desafios científicos e tecnológicos que não foram aqui detalhados, tais como a preservação da privacidade dos usuários, a criação de modelos de recomendação inter-domínios, o desenvolvimento de sistemas descentralizados operando em redes computacionais distribuídas, a otimização de sistemas para sequências de recomendações, a otimização de sistemas para dispositivos móveis e outros. Entretanto, esses desafios são menos relevantes ao nosso projeto porque não se aplicam diretamente à recomendação de produtos de e-commerces via e-mail, ou porque não se encaixam no formato dos dados que serão utilizados para análise.

3 Palavras-chave (Português-Inglês)

As palavras chave principais desse Trabalho de Conclusão de Curso são:

1. Sistema de recomendação (recommender system)
2. e-commerce (*idem*)
3. Recomendação baseada em conteúdo (content-based recommendation)
4. Filtragem colaborativa (collaborative filtering)
5. Recomendação híbrida (hybrid recommendation)

4 Procedimento de busca

A busca por artigos científicos começou com a recomendação de livros didáticos pelo professor orientador ([7] e [4]). O primeiro livro-texto aborda a temática com noções introdutórias, apresentando os conceitos básicos e algoritmos de filtragem colaborativa baseada em usuários e baseada em ítens, assim como algumas medidas de similaridade. O segundo livro didático, por sua vez, explora o tema de maneira bem mais detalhada, tratando de temas avançados, como algorítmos de aprendizado continuado e de sistemas robustos contra ataques (por exemplo, quando uma grande quantidade de perfis falsos são inseridos na plataforma). A partir da leitura dos capítulos introdutórios de ambas as fontes, observamos as palavras-chave ligadas à temática de Sistemas de Recomendação e iniciamos a pesquisa por artigos especializados.

No website Google Scholar buscamos as palavras-chave indicadas na Seção 3 e priorizamos os artigos mais citados, e principalmente os documentos de *surveys* e *reviews*.

Na base de dados IEEE Explorer também procuramos pelas palavras chave e selecionamos os artigos mais citados, mas filtramos por data a fim de obter apenas resultados recentes (artigos publicados após 2010).

Uma vez obtidos os documentos, avaliamos se o conteúdo de cada um deles era relevante ao escopo do trabalho. Como não dispomos de uma grande de dados demográficos dos clientes, descartamos artigos que se baseavam inteiramente em filtragem colaborativa baseada em usuários ou que requeriam dados de nevegação, tais como [8]. Em vez disso, priorizamos artigos que exploram técnicas de recomendação item-a-item ou mesmo técnicas híbridas, como [9], [10] e [11]. O primeiro dos textos, escrito por desenvolvedores e pesquisadores da Amazon, defende que a filtragem colaborativa baseada em itens é superior que a baseada em usuários em termos de escalabilidade e qualidade de recomendação, e serviu de norte para a nossa pesquisa de referências. O segundo artigo confirma essa argumentação e dá subsídios teóricos para a construção de algorítmos de filtragem e de medidas de similaridade. O terceiro, por sua vez, utiliza dados dos atributos dos itens e também das avaliações dos usuários para construir um modelo mais acurado, sustentando uma melhor recomendação.

5 Estado da arte do problema. Estado da arte das soluções

Do ponto de vista do estado da arte do problema, pode-se formulá-lo como se segue [5]:

Seja C o conjunto de todos os usuários e seja S o conjunto de todos os itens que podem ser recomendadas, tais como livros, filmes ou restaurantes. Seja u uma função de utilidade, que mede a utilidade do produto s ao usuário c, ou seja, $u : C \times S \rightarrow R$, onde R é um conjunto totalmente ordenado (por exemplo, números inteiros não-negativos ou números reais dentro de um determinado intervalo). Em seguida, para cada usuário $c \in C$, queremos escolher o item $s' \in S$ que maximize a utilidade do usuário. Mais formalmente:

$$\forall c \in C, \ s'_c = \arg\max_{s \in S} u(c, s)$$

Nesse problema, destacam-se três grupos de estratégias de sugestão de itens:

- Recomendações baseadas em conteúdo: o usuário recebe recomendações de itens semelhantes aos preferidos no passado;
- Recomendações colaborativas: o usuário recebe recomendações de itens que que pessoas com gostos e preferências semenlhantes gostaram no passado;
- Recomendações híbridas: esses métodos combinam métodos colaborativos e métodos baseados em conteúdo.

As técnicas de recomendação, por sua vez, dividem-se em dois grupos:

- Técnicas baseadas em heurísticas ou memória: essencialmente fazem a previsão das classificações com base em toda a coleção de itens anteriormente classificados pelos usuários – com técnicas tais como do vizinho mais próximo;
- Técnicas baseadas em modelos: utilizam o conjunto de avaliações com o objetivo de descrever um modelo – tais como redes Bayesianas –, que é então usado para fazer a previsão das classificações.

As técnicas de recomendação baseadas em conteúdo exploram os dados dos itens para calcular a sua relevância conforme o perfil do usuário. As pesquisas nesta área são realizadas em diversas frentes como a de Recuperação de Informação e Inteligência Artificial.

Os sistemas baseados em Recuperação de Informação partem do princípio que os usuários interessados em recomendações estão engajados em um processo de busca de informação. Nesses sistemas, o usuário fornece uma lista de palavras-chave para uma busca específica. Em sistemas baseados em filtragem de conteúdo, porém, a informação base para as recomendações é o perfil do usuário. Os itens a serem recomendados podem possuir diversos atributos e formas de classificação, cada um podendo ser descrito por uma pequena quantidade de atributos com valores conhecidos. Entretanto, documentos como emails, websites ou reviews de usuários são

compostos por textos sem estrutura definida e a abordagem por Recuperação de Informação é mais recomendada [12].

Na abordagem por Inteligência Artificial, a recomendação pode ser vista como um problema de aprendizado que explora os conhecimentos sobre o usuário. Na sua forma mais simples, os perfis de usuários estão especificados por palavras-chave que refletem os interesses de longo prazo dos usuários. Muitas vezes é recomendado que o aprendizado seja feito com base no perfil do usuário conforme o uso continuo, ao invés de forçá-lo a responder diversas perguntas demográficas [1].

Isso geralmente envolve problemas de aprendizado da máquina, onde o objetivo é aprender a categorizar novas informações baseadas em informações previamente adquiridas e rotuladas como interessantes ou não pelo usuário. Com estas informações em mão, metodos de aprendizado de máquina são capazes de gerar modelos preditivos que, com nova informação, recomendarão um item que tem mais chances de ser do interesse do consumidor.

Do ponto de vista do estado da arte das soluções, as variáveis de interesse estão ligadas do número de usuários no sistema, ao número de itens, ao nível de dispersão, à medida de qualidade da recomendação e ao custo computacional [13].

No que se refere à dependência do número de usuários, a filtragem colaborativa a base de usuários é extremamente efetiva para um baixo número de usuários, mas tem uma dependência quase constante em relação a essa quantidade. A filtragem colaborativa a base de itens é consideravelmente pior para um baixo número de usuários, mas supera todos os outros métodos baseados em memória para quantidades maiores.

A dependência do número de ítens é, de certa forma, oposta à de usuários: a filtragem colaborativa a base de itens é extremamente efetiva para poucos itens, mas tem uma dependência quase constante no número de itens. A filtragem colaborativa baseada em usuários tem performance consideravelmente pior de início, mas supera todos os outros métodos baseados em memória para maiores quantidades de usuários.

Com relação ao nível de dispersão dos dados, a filtragem baseada em usuários e a baseada em itens mostram uma dependência semelhante. Na medida de qualidade de recomendação (menor erro quadrático médio), todos os métodos de recomendação variam não-linearmente com o número de usuários, itens e nível de dispersão, e de modo geral há um trade-off entre a acurácia e o tempo de processamento da sugestão de produtos. Outros métodos de recomendação são também explorados no artigo [13], mas em virtude de serem menos tradicionais, não foram incluídos na presente análise.

Referências

[1] Kangning Wei, Jinghua Huang e Shaohong Fu. "A survey of e-commerce recommender systems". Em: *Service Systems and Service Management, 2007 International Conference on*. IEEE. 2007, pp. 1–5.

[2] Leslie Tutol. *Amazon Launches 'Login and Pay with Amazon' for a Seamless Buying Experience*. Novembro de 2013. URL: http://services.amazon.com/post/Tx2A98P3EKP62O2/Amazon-Launches-Login-and-Pay-with-Amazon-for-a-Seamless-Buying-Experience.

[3] Valentina Palladino. *Amazon sold 426 items per second in run-up to Christmas*. Dezembro de 2013. URL: http://www.theverge.com/2013/12/26/5245008/amazon-sees-prime-spike-in-2013-holiday-season.

[4] Pasquale Lops, Marco de Gemmis e Giovanni Semeraro. Em: *Recommender Systems Handbook*. Springer, 2011.

[5] Gediminas Adomavicius e Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions". Em: *Knowledge and Data Engineering, IEEE Transactions on* 17.6 (2005), pp. 734–749.

[6] J Ben Schafer, Joseph Konstan e John Riedl. "Recommender systems in e-commerce". Em: *Proceedings of the 1st ACM conference on Electronic commerce*. ACM. 1999, pp. 158–166.

[7] Toby Segaran. *Programming Collective Intelligence*. First. O'Reilly, 2007, pp. 7–28. ISBN: 9780596529321.

[8] Pooyan Adibi e Behrouz Tork Ladani. "A collaborative filtering recommender system based on user's time pattern activity". Em: *Information and Knowledge Technology (IKT), 2013 5th Conference on*. IEEE. 2013, pp. 252–257.

[9] Greg Linden, Brent Smith e Jeremy York. "Amazon.com recommendations: Item-to-item collaborative filtering". Em: *Internet Computing, IEEE* 7.1 (2003), pp. 76–80.

[10] Badrul Sarwar et al. "Item-based collaborative filtering recommendation algorithms". Em: *Proceedings of the 10th international conference on World Wide Web*. ACM. 2001, pp. 285–295.

[11] Sutheera Puntheeranurak e Thanut Chaiwitooanukool. "An Item-based collaborative filtering method using Item-based hybrid similarity". Em: *Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference on*. IEEE. 2011, pp. 469–472.

[12] J Ben Schafer, Joseph A Konstan e John Riedl. "E-Commerce Recommendation Applications". Em: *Data Mining and Knowledge Discovery* 5 (2001), pp. 115–153.

[13] Joonseok Lee, Mingxuan Sun e Guy Lebanon. "A comparative study of collaborative filtering algorithms". Em: *arXiv preprint arXiv:1205.3193* (2012).