Escola Politécnica da Universidade de São Paulo

PMR2500 – Projeto de Conclusão do Curso I

Estado da Arte

Desenvolvimento de um Sistema de Recomendação para e-Commerces

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

Orientador

Prof. Dr. Fábio Gagliardi Cozman

6 de junho de 2014

1 Estado da Arte

As terminologias *cliente* e *usuário* neste texto serão intercambiáveis e sem distinção semântica, mesmo que na prática essas duas entidades possam ser diferentes. Da mesma forma, *item* e *produto* terão o mesmo significado neste trabalho.

A fim de tornar a formulação mais genérica, também não faremos distinção entre *avaliação positiva* de um item e *compra* de um item. Avaliação positiva é toda avaliação r_{ui} do item i feital pelo usuário u tal que $r_{ui} > M$, e avaliação negativa tal que $r_{ui} \leq M$, sendo M um valor mínimo escolhido a priori, indicador de que o usuário u "gostou" do item i. No caso de um banco de dados sem avaliações dos produtos, será levada em conta a compra dos itens e será admitida avaliação unitária e valor mínimo nulo. Desta forma, os bancos de dados que contenham informações do tipo "usuário u avaliou o item i em $r_{ui} = 3.54 > M$" e aqueles que contenham "usuário u comprou o item i, logo $r_{ui} = 1 > 0$" serão tratados equivalentemente.

1.1 Estado da arte dos problemas

O problema de recomendação pode ser formulado como se segue, adaptado da referência (ADOMAVICIUS; TUZHILIN, 2005), com notação inspirada em (SYMEONIDIS; NANOPOULOS; MANOLOPOULOS, 2007):

"Seja $\mathcal{U}$ o conjunto de todos os usuários e seja $\mathcal{I}$ o conjunto de todos os itens que podem ser recomendados, tais como livros, filmes ou artigos científicos. Seja ℓ uma função de utilidade, que mede a relevância do produto i para usuário u, ou seja, $\ell : \mathcal{U} \times \mathcal{I} \to \mathcal{R}$, onde $\mathcal{R}$ é um conjunto totalmente ordenado – por exemplo, números inteiros ou números reais dentro de um determinado intervalo, em geral $\{-1, 0, +1\}$ ou $[1, 5]$. O objetivo do sistema de recomendação é determinar o item $\tilde{\imath}_u$ que maximize a utilidade ℓ_{ui} do usuário u."

$$\forall u \in \mathcal{U},\ \tilde{\imath}_u = \arg\max_{i \in \mathcal{I}} \ell_{ui} \tag{1.1}$$

O problema central da recomendação é que "em geral a função ℓ é desconhecida ou não é definida para todo o espaço $\mathcal{U} \times \mathcal{I}$", e portanto determinar $\tilde{\imath}$ através da equação 1.1 é inviável.

Em algumas formulações, "a utilidade é descrita pela avaliação r_{ui} do item i feita pelo usuário u". Neste caso, o sistema de recomendação busca determinar $\hat{r}_{ui}$ que melhor se aproxime de r_{ui}, e a qualidade da recomendação é normalmente descrita pela distância

entre esses dois valores. Em outros sistemas, todavia, a utilidade é descrita diferentemente, não bastando que um item tenha o maior valor de $\hat{r}_{ui}$ para que ele seja recomendado.

Para lidar o problema da recomendação, existem três grandes grupos de estratégias de sugestão de itens, segundo as referências (ADOMAVICIUS; TUZHILIN, 2005; BALABANOVIC; SHOHAM, 1997):

- Recomendações baseadas em conteúdo: o usuário recebe sugestões de itens similares àqueles pelos quais ele se interessou no passado;
- Recomendações colaborativas: o usuário recebe sugestões de itens que pessoas com preferências semelhantes gostaram no passado;
- Recomendações híbridas: esses métodos combinam caracteríticas de sistemas colaborativos e baseados em conteúdo. O usuário recebe sugestões de itens compatíveis com seu perfil e itens do interesse de usuários com perfil similar.

As estratégias de recomendação baseadas em conteúdo exploram os dados dos itens para calcular a sua relevância conforme o perfil do usuário. Suas técnicas de recomendação podem ser classificadas em dois grupos, aquelas baseadas em heurísticas ou memória – essencialmente fazem a previsão com base em toda a coleção de itens anteriormente classificados pelos usuários – e aquelas baseadas em modelos – utilizam o conjunto de avaliações com o objetivo de descrever um modelo, como em uma regressão linear ou em uma rede Bayesiana.

Em sistemas baseados em conteúdo, os itens a serem recomendados podem possuir diversos atributos e formas de classificação. Em documentos como e-mails, websites ou reviews de usuários, os itens são textos sem estrutura definida e a abordagem mais comum é a de recuperação de informação – o usuário procura por uma lista de termos desejados e o sistema retorna os textos que contém aqueles termos com maior relevância, tal como é feito em um motor de busca (SCHAFER; KONSTAN; RIEDL, 2001). Nesses casos, calcula-se a similaridade entre documentos a partir de formulações que levam em conta as palavras ou termos escritos, como a TF-IDF ou o classificador Bayesiano (LOPS; GEMMIS; SEMERARO, 2011b).

Na abordagem de sistemas baseados em conteúdo, a recomendação pode ser vista como um problema de aprendizado de máquina, em que o sistema adquire conhecimento sobre o usuário. Muitas vezes é recomendado que o aprendizado seja feito com base no perfil do usuário em uso continuo, ao invés de forçá-lo a responder diversas perguntas demográficas (WEI; HUANG; FU, 2007) – idade, gênero, classe social, etc. O objetivo é categorizar novas informações baseadas em informações previamente adquiridas e rotuladas como interessantes ou não pelo usuário. Com estas informações em mão, é possível gerar modelos preditivos que evoluem conforme aparecem novas informações.

As recomendações colaborativas, por sua vez, tentam prever a utilidade dos itens para cada cliente baseado em itens previamente avaliados por outros usuários. Elas podem ser baseadas em usuários, isto é, na escolha de clientes que possuam avaliações similares de produtos, quanto baseadas em itens, na escolha de produtos avaliados similarmente (LINDEN; SMITH; YORK, 2003). Admite-se que a filtragem colaborativa é baseada em usuários, caso não seja especificado o contrário.

Mais formalmente, quando baseada em usuários, a utilidade ℓ_{ui} de um item i para um usuário u é estimada com base nas utilidades $\ell_{v_k^u i}$ dos usuários $v_k^u \in \mathcal{U}$ que são "similares" ao usuário u. De maneira análoga, quando baseada em itens, a utilidade ℓ_{ui} é prevista com base nas utilidades $\ell_{uj_k^u}$, dado itens $j_k^u \in \mathcal{I}$ que são "similares" aos itens i .

Por fim, as recomendações híbridas combinam aspectos tanto da filtragem colaborativa (baseada em usuários ou em itens) quanto da filtragem baseada em conteúdo, com o objetivo de atingir uma melhor recomendação ou de superar problemas recorrentes nas técnicas individuais, como a esparsidade (*sparsity*) dos dados ou o *cold start* (BURKE, 2007).

1.2 Estado da arte das soluções

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, à acurácia, à medida de qualidade da recomendação e ao custo computacional (LEE; SUN; LEBANON, 2012). Essa referência faz uma análise extensiva de diversos métodos de filtragem colaborativa.

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 itens 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 à acurácia 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 acurácia, e de modo geral há um *trade-off* entre a dispersão dos dados e o tempo de processamento da sugestão de produtos.

1.3 Desafios científicos e tecnológicos

Um dos maiores desafios tecnológicos dos sistemas de recomendação é, atualmente, o da escalabilidade (WEI; HUANG; FU, 2007). O sistema de recomendação deverá ser flexível no sentido de poder operar igualmente bem tanto em conjuntos pequenos quanto em grandes bases de dados, que podem chegar até centenas de milhões de clientes (TUTOL, 2013) e de produtos (PALLADINO, 2013). Isso significa que as recomendações devem ser suficientemente rápidas e ainda assim prover sugestões valiosas aos consumidores.

Outra grande dificuldade é a esparsidade ou *sparsity* dos dados, ou seja, o fato de a maioria dos clientes nunca ter interagido com mais de algumas unidades de itens, fazendo com que a matriz de relação usuário-item tenha uma quantidade muito pequena de valores preenchidos, da ordem de 1% (FENNELL, 2009).

Um problema muito comum nos sistemas de recomendação é o do *cold start*: quando itens ou usuários são inicialmente introduzidos no sistema, existe pouca ou nenhuma informação sobre eles. O sistema é incapaz de realizar inferências sobre quais itens recomendar ao novo usuário ou sobre quais produtos são similares ao novo item.

Outro desafio científico é referente à diversidade das recomendações realizadas, também chamado de excesso de especialização ou *over-specialization* (ADOMAVICIUS; TUZHILIN, 2005). 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 as preferências de curto e longo termo dos usuários (SCHAFER; KONSTAN; RIEDL, 1999). 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 (WEI; HUANG; FU, 2007).

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. Um sistema de recomendação inteligente também deveria prever quando enviar uma determinada recomendação, e não agir apenas mediante requisição dos clientes (LOPS; GEMMIS; SEMERARO, 2011a). Entretanto, esses desafios são menos relevantes porque não se aplicam diretamente aos objetivos do nosso projeto, que serão especificados no relatório final.

Referências

ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, IEEE, v. 17, n. 6, p. 734–749, 2005. Citado 3 vezes nas páginas 2, 3 e 5.

BALABANOVIC, M.; SHOHAM, Y. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, v. 40, p. 66–72, 1997. Citado na página 3.

BURKE, R. Hybrid web recommender systems. In: *The adaptive web*. [S.l.]: Springer, 2007. p. 377–408. Citado na página 4.

FENNELL, J. Collaborative filtering on sparse rating data for yelp. com. 2009. Citado na página 5.

LEE, J.; SUN, M.; LEBANON, G. A comparative study of collaborative filtering algorithms. *arXiv preprint arXiv:1205.3193*, 2012. Citado na página 4.

LINDEN, G.; SMITH, B.; YORK, J. Amazon.com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, IEEE, v. 7, n. 1, p. 76–80, 2003. Citado na página 4.

LOPS, P.; GEMMIS, M. de; SEMERARO, G. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. Citado na página 5.

LOPS, P.; GEMMIS, M. de; SEMERARO, G. Content-based recommender systems: State of the art and trends. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 73–105. Citado na página 3.

PALLADINO, V. *Amazon sold 426 items per second in run-up to Christmas*. 2013. Disponível em: <http://www.theverge.com/2013/12/26/5245008/amazon-sees-prime-spike-in-2013-holiday-season>. Citado na página 5.

SCHAFER, J. B.; KONSTAN, J.; RIEDL, J. Recommender systems in e-commerce. In: ACM. *Proceedings of the 1st ACM conference on Electronic commerce*. [S.l.], 1999. p. 158–166. Citado na página 5.

SCHAFER, J. B.; KONSTAN, J. A.; RIEDL, J. E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, v. 5, p. 115–153, 2001. Citado na página 3.

SYMEONIDIS, P.; NANOPOULOS, A.; MANOLOPOULOS, Y. Feature-weighted user model for recommender systems. In: *User Modeling 2007*. [S.l.]: Springer, 2007. p. 97–106. Citado na página 2.

TUTOL, L. *Amazon Launches 'Login and Pay with Amazon' for a Seamless Buying Experience*. 2013. Disponível em: <http://services.amazon.com/post/Tx2A98P3EKP62O2/Amazon-Launches-Login-and-Pay-with-Amazon-for-a-Seamless-Buying-Experience>. Citado na página 5.

WEI, K.; HUANG, J.; FU, S. A survey of e-commerce recommender systems. In: IEEE. *Service Systems and Service Management, 2007 International Conference on*. [S.l.], 2007. p. 1–5. Citado 2 vezes nas páginas 3 e 5.