

DESENVOLVIMENTO DE UMA BIBLIOTECA COMPUTACIONAL PARA SISTEMAS DE RECOMENDAÇÃO DE LOJAS DE COMÉRCIO ONLINE

Escola Politécnica da Universidade de São Paulo

Antônio Viggiano

`agfviggiano@gmail.com`

Fernando Fochi

`fernando.fochi@gmail.com`

Prof. Dr. Fábio Gagliardi Cozman

18 de novembro de 2014

Sumário

- 1 Introdução
- 2 Objetivos
- 3 Estado da Arte
- 4 Metodologia
- 5 Requisitos
- 6 Síntese de Soluções
- 7 Avaliação de Desempenho
- 8 Resultados
- 9 Cronograma

Introdução

Importância econômica

Annual B2C e-commerce sales in the United States from 2002 to 2013 (in billion U.S. dollars)

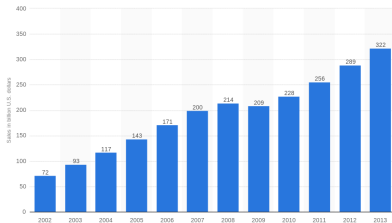
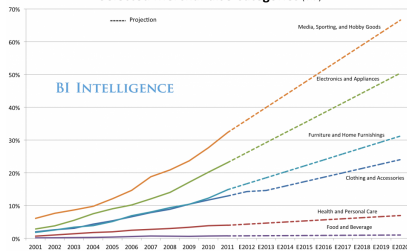


Figura 1: Vendas de varejo atribuídas a lojas online nos EUA (STATISTA, 2014)

Percent Of Retail Sales Attributable To Online In Selected Merchandise Categories (U.S.)



Source: U.S. Census, Internet Retailer, BI Intelligence Estimates

Figura 2: Percentual de vendas de varejo atribuídas a lojas online nos EUA por categoria (SMITH, 2014)

Introdução

Aplicação

The Facebook logo, consisting of the word "facebook" in white lowercase letters on a blue rectangular background.

Relações de amizade

Músicas

The Amazon.com logo, featuring the text "amazon.com" in black with a yellow curved arrow underneath the word "amazon".

Livros **35 %**
(MARSHALL, 2006)

The Google News logo, with "Google" in its multi-colored font and "news" in blue.

Notícias **38 %**
(DAS et al., 2007)

Filmes **75 %**
(AMATRIAIN, 2012)

Introdução

O que são Sistemas de Recomendação?

Definição

“São ferramentas e técnicas de software destinadas a prover sugestões de itens para usuários” (RICCI; SHAPIRA, 2011)

Etapas principais

- Aquisição dos dados de entrada
- Determinação das recomendações
- Apresentação dos resultados ao usuário

Objetivos

- **Sistema de recomendação** de produtos para **e-commerces**
 - Propostas de diferentes algoritmos
- **Análise de desempenho** das recomendações
 - Validação cruzada
 - **Acurácia e Precisão**



m o v i e l e n s
helping you find the *right* movies

Estado da Arte

Problema

\mathcal{U} Conjunto dos usuários u

\mathcal{I} Conjunto dos itens i

r_{ui} Histórico avaliações

ℓ Função de utilidade

• $\ell : \mathcal{U} \times \mathcal{I} \rightarrow \mathcal{R}$ p.ex. $\{-1, 0, +1\}$ ou $[1, 5]$

Objetivo

Determinar o item \tilde{i}_u que maximize a utilidade ℓ_{ui} do usuário u :

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

Problema

ℓ desconhecida

Estado da Arte

Soluções

Estratégias de recomendação

- Colaborativas
- Conteúdo
- Híbridas

Utilização comercial (CHIANG, 2012)

- Netflix** Filtragem colaborativa
- Amazon** Filtragem baseada em conteúdo
- Pandora** Experts + votos positivos/negativos
- YouTube** Contagem de visitas mútuas

Estado da Arte

Soluções

Filtragem colaborativa (CF)

- Usuário-usuário
- Item-item

Filtragem de conteúdo (CB)

Métodos híbridos (H)

- CF + CB

Tabela 1: Avaliações r_{ui}

	i_1	i_2	i_3	i_4
u_1	-	4	3	-
u_2	-	4	3	5
u_3	2	5	-	1

Tabela 2: Atributos a_{if}

	f_1	f_2	f_3	f_4
i_1	1	50	0.8	P
i_2	0	75	0.3	M
i_3	1	30	0.4	G

Estado da Arte

Desafios

Filtragem colaborativa (CF)

- *Cold start*
- Esparsidade

Filtragem de conteúdo (CB)

- Excesso de especialização
- Análise “superficial” do conteúdo

Todos os métodos (CF, CB, H)

- Escalabilidade

Tabela 3: Avaliações r_{ui}

	i_1	i_2	\dots	i_{100}
u_1	-	4	\dots	-
u_2	-	2	\dots	-
u_3	5	-	\dots	-

Tabela 4: Atributos a_{if}

	f_1	f_2	f_3
i_1	1	50	0.8
i_2	1	50	0.8
i_3	0	75	0.3

Metodologia

- Definição de **necessidades**
- Definição dos **parâmetros de sucesso**
- Proposição de **soluções**
- Implementação e testes em **bancos de dados simples**
- Teste em **bancos de dados reais**

Requisitos

Requisitos funcionais

- EMA máximo:
- 20% para Precisão
- 20% para Abrangência
- *Throughput* mínimo
 - 100 mil recomendações por hora

Requisitos não funcionais

- Escalabilidade
- Sistema genérico
 - Padronização dos dados de entrada/saída
- Código aberto

Síntese de Soluções

Ponderação de Atributos

$$s_{ij} = \sum_f w_f (1 - d_{fij})$$

Perfil de Usuários

$$S_{uv} = \frac{\sum_{f \in \mathcal{F}_{uv}} w_{uf} w_{vf}}{\sqrt{\sum_{f \in \mathcal{F}_{uv}} w_{uf}^2} \sqrt{\sum_{f \in \mathcal{F}_{uv}} w_{vf}^2}}$$

Tabela 5: Medidas de distância entre alguns atributos

Atributo f	Domínio F	Distância d_f
Marca	Literal	$1 - \delta_{ij}^f$
Cor	$(\mathbb{N} \setminus \mathbb{N}_{256})^3$	$\frac{\ a_{if} - a_{jf}\ _2}{\max_{i,j} \ a_{if} - a_{jf}\ _2}$
Preço	\mathbb{R}	$\frac{ a_{if} - a_{jf} }{\max_{i,j} a_{if} - a_{jf} }$

Avaliação de Desempenho

- Distância entre recomendações
 - $EMA = |\hat{i} - i|$
- Desempenho mediante a mudança nas variáveis
 - Quantidade de atributos utilizados
- Tempo de execução
 - Em função do algoritmo
 - Em função do tamanho do banco de dados

Tabela 6: Avaliação de sistemas de predição

Medida	Fórmula	Significado
Precisão	$\frac{VP}{VP+FP}$	% Predições corretas de casos positivos
Acurácia	$\frac{VP+VN}{VP+VN+FP+FN}$	% Predições corretas

Resultados

Primeiros testes

Pesos unitários

$$s_{ij} = \sum_f (1 - d_{fij})$$

13 s Tempo de inicialização para $|\mathcal{R}| = 100$ mil

8 min Cálculo de s_{ij} para $|\mathcal{I}| = 1000$

100% CPU
2.80GHz \times 4

420 MB Memória

60 dias Para $|\mathcal{I}| = 100$ mil

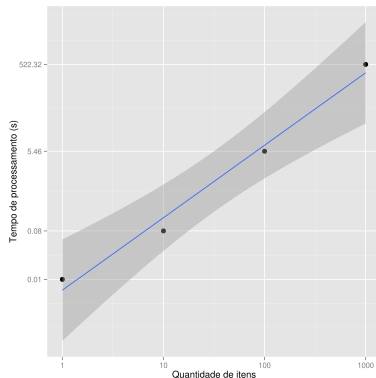


Figura 3: Tempo de processamento em função do número de itens em $\mathcal{O}(n^2)$

Resultados

Aquisição de dados

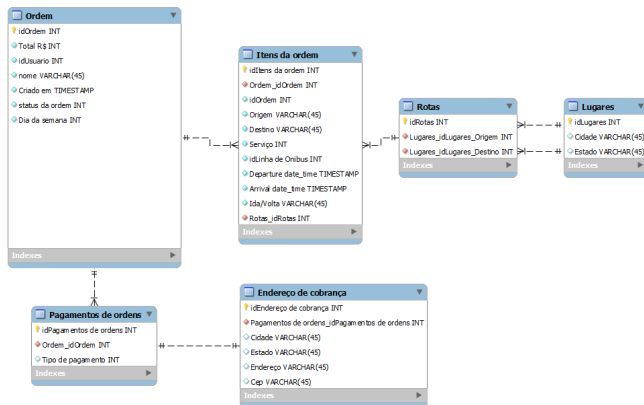


Figura 4: Banco de dados de um e-commerce de passagens de ônibus

Cronograma

- 09/07 Pré-tratamento do banco de dados
- 16/07 **Programação** do método Ponderação de Atributos
- 23/07 **Programação** do método Perfil de Usuários
- 30/07 Análise comparativa dos dois algoritmos

- 13/08 Relatório de atividades de implementação
- 27/08 Primeiros testes com o sistema
(**precisão e acurácia** para uma base de testes)

- 03/09 Testes com o sistema (**validação cruzada**)
- 24/09 Melhorias incrementais e relatório de atividades

- 15/10 **Relatório aprofundado** de atividades

- 05/11 Elaboração da apresentação e finalização dos relatórios
- 12/11 Melhorias incrementais

Bibliografia I

- ▶ AMATRIAIN, X. *Netflix Recommendations: Beyond the 5 stars*. 2012. Disponível em: <<http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html>>.
- ▶ CHIANG, M. *Networked Life: 20 Questions and Answers*. Cambridge University Press, 2012. (BusinessPro collection). ISBN 9781107024946. Disponível em: <<http://books.google.com.br/books?id=N5DJJXoLPDQC>>.
- ▶ DAS, A. S. et al. Google news personalization: scalable online collaborative filtering. In: ACM. *Proceedings of the 16th international conference on World Wide Web*. [S.l.], 2007. p. 271–280.

Bibliografia II

- ▶ MARSHALL, M. *Aggregate Knowledge raises \$5M from Kleiner, on a roll*. 2006. Disponível em: <<http://venturebeat.com/2006/12/10/aggregate-knowledge-raises-5m-from-kleiner-on-a-roll/>>.
- ▶ RICCI, L. R. F.; SHAPIRA, B. Introduction to recommender systems handbook. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 1–35.
- ▶ SMITH, C. *E-COMMERCE AND THE FUTURE OF RETAIL: 2014 [SLIDE DECK]*. 2014. Disponível em: <http://www.businessinsider.com/the-future-of-retail-2014-slide-deck-sai-2014-3?nr_email_referer=1&utm_source=Triggermail&utm_medium=email&utm_content=emailshare>.

Bibliografia III

- ▶ STATISTA. *Annual B2C e-commerce sales in the United States 2002-2013*. 2014. Disponível em: <<http://www.statista.com/statistics/271449/annual-b2c-e-commerce-sales-in-the-united-states/>>.