

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

Sumário

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

Introdução

Importância econômica

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

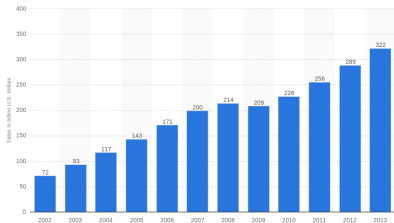
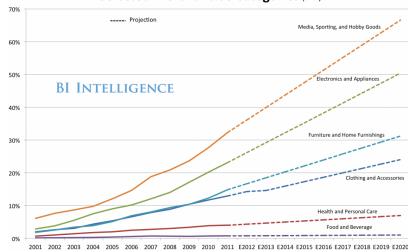


Figura 1: Vendas de varejo atribuídas a lojas online nos EUA (STATISTA, 2014)

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



facebook.

Relações de amizade



lost.fm

Músicas



amazon.com

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



Google news

Notícias **38 %**
(DAS et al., 2007)



NETFLIX

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

Objetivos

- **Biblioteca computacional para sistemas de recomendação**
 - Abrangente e adaptável
 - Leitura de dados e cálculo de sugestões
- **Análise de desempenho**
 - Validação cruzada
 - Precisão e Abrangência



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

Sistemas de recomendação

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



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

Utilização comercial (CHIANG, 2012)

Estratégias de recomendação

- Colaborativas
- Conteúdo
- Híbridas

- 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

Requisitos

- 20% Precisão
- 20% Abrangência

Tabela 3: Avaliação de sistemas de predição

Medida	Fórmula	Significado
Precisão	$\frac{VP}{VP+FP}$	Porcentagem de casos positivos corretamente preditos.
Abrangência	$\frac{VP}{VP+FN}$	Porcentagem de casos positivos sobre aqueles que foram marcados como positivos.
F_1	$2 \cdot \frac{\text{Precisão} \cdot \text{Abrangência}}{\text{Precisão} + \text{Abrangência}}$	Média harmônica entre precisão e abrangência.

Metodologia

Estruturação do banco de dados

100k 100 000 avaliações de 943 usuários
para 1682 filmes

IMDB 28 819 filmes

IMDB-100k 943 usuários, 1682 filmes e 25 atributos

Metodologia

Desenvolvimento da biblioteca

Ferramenta utilizada

RStudio Editor de texto e console

Estrutura da biblioteca

```
recsys/  
|-- db  
|   |-- ml-100k  
|       |-- u.data  
|       |-- u.item  
|       |-- u.user  
|       |-- ...  
|-- methods  
|   |-- fw.R  
|   |-- ui.R  
|   |-- up.R  
|-- results  
|   |-- benchmark.R  
|   |-- performance.R  
|   |-- run_tests.R  
|-- setup  
|   |-- functions.R  
|   |-- setup.R
```

Metodologia

Validação cruzada

Ambiente de testes

- Máquina r3.large
- 2 vCPU
- 15 GB de memória RAM
- Amazon Linux AMI release 2014.09 x86_64
- Custo total R\$ 5,70

Avaliação

- $T = 75\%$ base de treinamento
- $H = 75\%$ dados “escondidos”

Tabela 4: Avaliações r_{ui}

	i_1	i_2	i_3	i_4
u_1	-	4	3	5
u_2	2	5	-	1
u_3	3	-	-	2
u_4	(5)	(2)	(3)	4

Síntese de Soluções

Ponderação de Atributos (FW)

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

Perfil de Usuários (UP)

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

Perfil Usuário-Item (UI)

$$\omega_{ui} = \sum_f w_{uf} a_{if}$$

Resultados

Tabela 5: Parâmetros de influência no desempenho dos algoritmos de recomendação

Variável	Descrição	Valor padrão
N	Lista de recomendação	20
T	Base de treinamento	75%
H	Avaliações “escondidas”	75%
M	Avaliações positivas	2
k	Vizinhos mais próximos	10
\mathcal{F}	Conjunto de atributos dos itens	Todos atributos
d^f	Medida de distância	Distância $L_1 \cdot ^f$
W	Quantidade de pesos	Todo $w_f > 0$

Resultados

Tamanho da lista de recomendação N

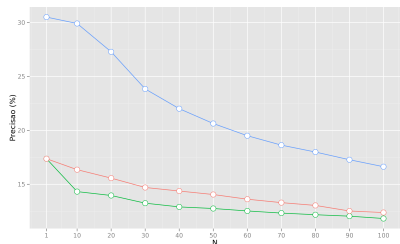


Figura 3: Precisão $\times N$

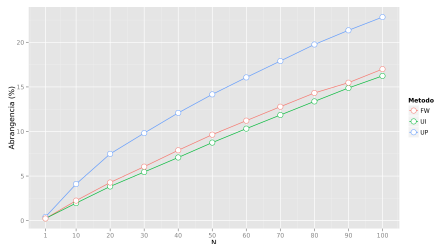


Figura 4: Abrangência $\times N$

Resultados

Tamanho da lista de recomendação N

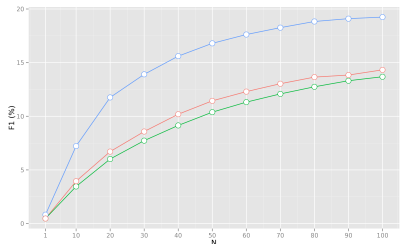


Figura 5: $F_1 \times N$

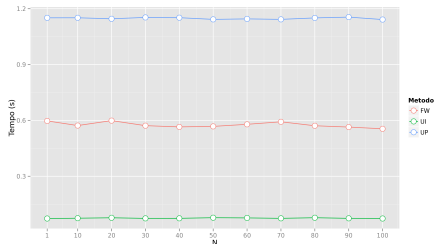


Figura 6: Tempo $\times N$

Resultados

Percentual da base de aprendizado T

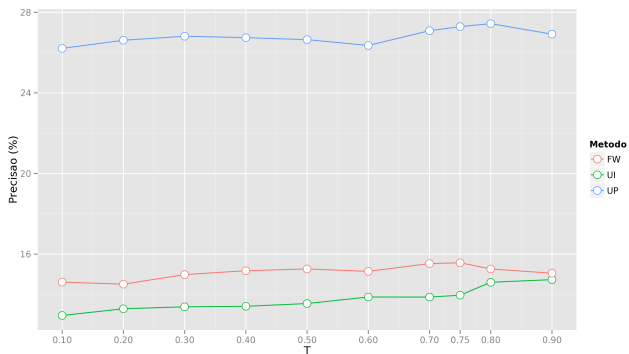


Figura 7: Precisão $\times T$

Abrangência, F_1 e Tempo praticamente constantes

Resultados

Percentual de avaliações “escondidas” dos usuários-teste H

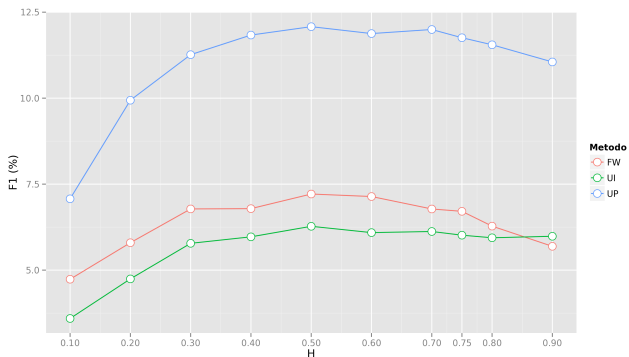


Figura 8: $F_1 \times H$

Precisão cresce e Abrangência decresce

Resultados

Valor mínimo para avaliações positivas M

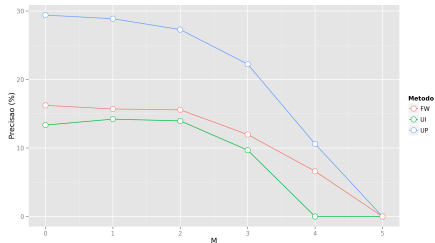


Figura 9: Precisão $\times M$

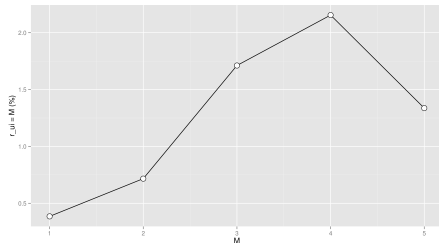


Figura 10: $\{\% r_{ui} = M\} \times M$

Resultados

Número de vizinhos mais próximos k

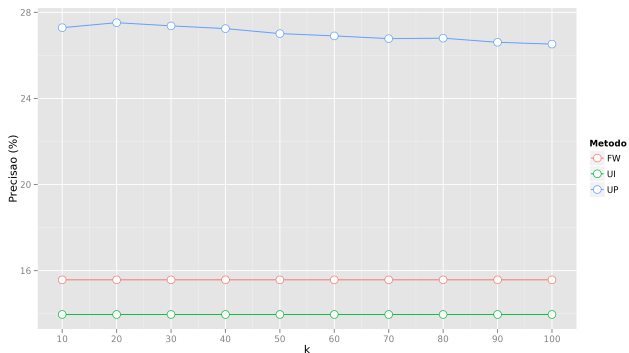


Figura 11: Precisão $\times k$

Precisão e Abrangência máximas para $k = 20$

Resultados

Conjunto de atributos dos itens \mathcal{F}

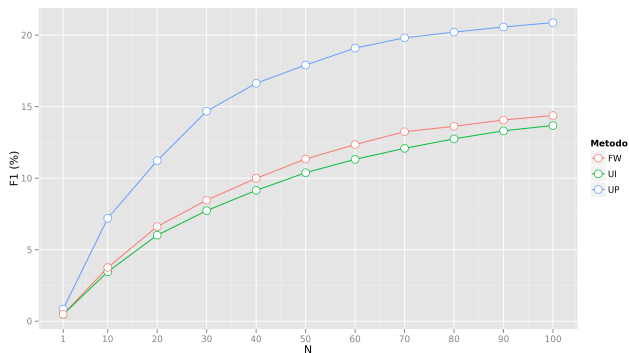


Figura 12: $F_1 \times \mathcal{F}$

Precisão e Abrangência aumentam com remoção dos atributos
{data de lançamento, ano}

Resultados

Medida de distância entre atributos d^f

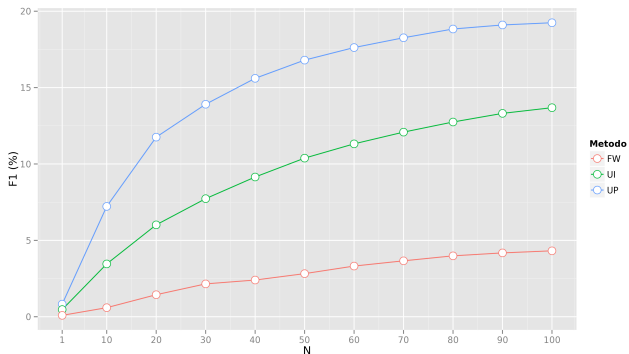


Figura 13: $F_1 \times$ diferentes d^f

Precisão e Abrangência diminuem com $d^f = J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

Resultados

Quantidade de pesos dos atributos W

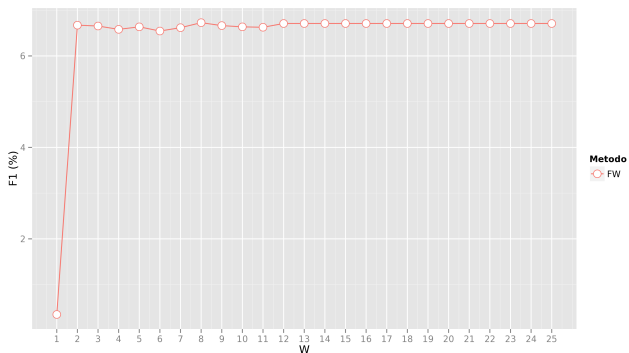


Figura 14: $F_1 \times W$

Precisão e Abrangência praticamente constantes

Conclusão

Discussão

- Dependência entre qualidade de recomendação e N
- Mais avaliações H é melhor que mais usuários T
- Muita influência de d^f , \mathcal{F}
- Pouca influência de k , M , W

Trabalhos futuros

- “Sistema de Recomendação nas Nuvens”
- Eliminação de restrições de entrada/saída de dados
- Desenvolvimento de um *driver* SQL
- Reconstrução da biblioteca em C
- Aplicação em um banco de dados de um e-commerce real

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/>>.