# Aplicação de Estratégias de Recomendação ao Domínio de Componentes de Software

#### Tássia Camões Araújo

Universidade de São Paulo

EXAME DE QUALIFICAÇÃO DE MESTRADO

Programa: Ciência da Computação Orientador: Prof. Dr. Arnaldo Mandel

24 de fevereiro de 2011



#### Roteiro

- 🚺 Introdução
- Sistemas de Recomendação
- Distribuições GNU/Linux
- Recomendação nas distribuições
- Desenvolvimento do Trabalho
- Conclusão

- Panorama atual: consumidores produtores
- Desenvolvimento software livre
- Importância de recomendações ("mais é menos")
- Sistemas recomendadores



- Panorama atual: consumidores produtores
- Desenvolvimento software livre
- Importância de recomendações ("mais é menos")
- Sistemas recomendadores



- Panorama atual: consumidores produtores
- Desenvolvimento software livre
- Importância de recomendações ("mais é menos")
- Sistemas recomendadores



- Panorama atual: consumidores produtores
- Desenvolvimento software livre
- Importância de recomendações ("mais é menos")
- Sistemas recomendadores



# Sistemas de Recomendação

- Ações
- Desafios
- Técnicas
- Estratégias
- Avaliação

# Sistemas de Recomendação

- Ações
- Desafios
- Técnicas
- Estratégias
- Avaliação

# Sistemas de Recomendação

- Ações
- Desafios
- Técnicas
- Estratégias
- Avaliação

- K-NN
- Bayes ingênuo
- Tf-idf
- BM25
- Apriori



- K-NN
- Bayes ingênuo
- Tf-idf
- BM25
- Apriori

$$c_{MAP} = \underset{c \in C}{\operatorname{arg\ max}} \ \hat{P}(c) \prod_{1 < i < n} \hat{P}(x_i | c)$$

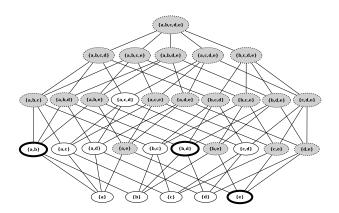
- K-NN
- Bayes ingênuo
- Tf-idf
- BM25
- Apriori

$$tf$$
- $idf_{t,d} = tf_{t,d} \cdot \log \frac{N}{df_t}$ 

- K-NN
- Bayes ingênuo
- Tf-idf
- BM25
- Apriori

$$R_{d,q} = \sum_{t \in q} \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)} \cdot \frac{tf_{t,D}(k_1+1)}{k_1((1-b)+b\frac{l_d}{l_{avg}})+tf_{t,D}} \cdot \frac{(k_3+1)qtf_{t,Q}}{k_3+qtf_{t,Q}}$$

- K-NN
- Bayes ingênuo
- Tf-idf
- BM25
- Apriori



#### Reputação dos itens

- Baseada IMDb > Laranja Mecânica (1971) > IMDb user reviews
- Golabora
- Baseada
- Baseada
- Híbrida





Update Data

IMDb user reviews for

#### Laranja Mecânica (1971) More at IMDbPro »

A Clockwork Orange (original title)
Filter: Best ▼ Hide Spoilers: □

Page 1 of 112: [1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] >

Index 1120 reviews in total

360 out of 469 people found the following review useful:

One of those films you have to watch multiple times, 28 April 2002

\*\*\*\*\*\*\*

Author: Agent 10 from Tucson, AZ

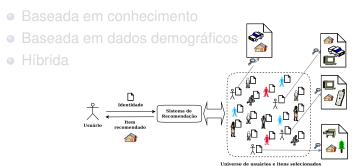
It's hard to judge a film such as this. Its cold and hard, yet can be exhilarating and sarcastic. It can be average, yet it can be visionary. Exploitive? Satirical? Too many questions to consider when one watches this film.

Even after 34 years, this film still speaks volumes about our current culture, which many ideals are ringing true today. The younger generations are out of control due to lack of parental control, lunk culture is becoming commonplace, violence is desensitizing the

- Reputação dos itens
- Baseada em conteúdo
- Colaborativa
- Baseada em conhecimento
- Baseada em dados demográficos
- Híbrida



- Reputação dos itens
- Baseada em conteúdo
- Colaborativa



- Reputação dos itens
- Baseada em conteúdo.
- Colaborativa
- Baseada em conhecimento

#### Customers Who Bought This Item Also Bought





\$7.68



Conjectures and Refutations: The Growth of Scienti ... by Karl Popper \*\*\*\* (11) \$16,42



Against Method (Fourth Edition) by Paul Feyerabend \*\*\*\*\* (21) \$14.49



The Open Society and Its Enemies, Vol.... by Sir Karl Raimund Popper \*\*\*\*\* (38)

\$19.77

- Reputação dos itens
- Baseada em conteúdo
- Colaborativa
- Baseada em conhecimento
- Baseada em dados demográficos
- Híbrida

- Reputação dos itens
- Baseada em conteúdo
- Colaborativa
- Baseada em conhecimento
- Baseada em dados demográficos
- Híbrida
  - ponderação, revezamento, combinação, combinação de atributos, cascata, acréscimo de atributos e meta-nível

- Reputação dos itens
- Baseada em conteúdo
- Colaborativa
- Baseada em conhecimento
- Baseada em dados demográficos
- Híbrida
  - ponderação, revezamento, combinação, combinação de atributos, cascata, acréscimo de atributos e meta-nível

#### Seleção dos dados

- Análise offline vs experimentos "ao vivo"
- Dados reais vs sintéticos

#### Métricas

- Acurácia de classificação
- Acurácia de predição
- Cobertura, curva de aprendizado, novidade, surpresa

- Seleção dos dados
  - Análise offline vs experimentos "ao vivo"
  - Dados reais vs sintéticos
- Métricas
  - Acurácia de classificação
  - Acurácia de predição
  - Cobertura, curva de aprendizado, novidade, surpresa

- Seleção dos dados
  - Análise offline vs experimentos "ao vivo"
  - Dados reais vs sintéticos
- Métricas
  - Acurácia de classificação
  - Acurácia de predição
  - Cobertura, curva de aprendizado, novidade, surpresa

- Seleção dos dados
  - Análise offline vs experimentos "ao vivo"
  - Dados reais vs sintéticos
- Métricas
  - Acurácia de classificação
  - Acurácia de predição
  - Cobertura, curva de aprendizado, novidade, surpresa

- Seleção dos dados
  - Análise offline vs experimentos "ao vivo"
  - Dados reais vs sintéticos
- Métricas
  - Acurácia de classificação
  - Acurácia de predição
  - Cobertura, curva de aprendizado, novidade, surpresa

- Seleção dos dados
  - Análise offline vs experimentos "ao vivo"
  - Dados reais vs sintéticos
- Métricas
  - Acurácia de classificação
  - Acurácia de predição
  - Cobertura, curva de aprendizado, novidade, surpresa

- Seleção dos dados
  - Análise offline vs experimentos "ao vivo"
  - Dados reais vs sintéticos
- Métricas
  - Acurácia de classificação
  - Acurácia de predição
  - Cobertura, curva de aprendizado, novidade, surpresa

- Aplicativos GNU + kernel Linux
- Seleção e configuração de softwares
- Escolha do Debian GNU/Linux como plataforma
  - Esquema consistente de distribuição de aplicativos: APT
  - Disponibilidade de dados estatísticos: Popcon
  - Possibilidade de integração dos resultados
  - Popularidade

- Aplicativos GNU + kernel Linux
- Seleção e configuração de softwares
- Escolha do Debian GNU/Linux como plataforma
  - Esquema consistente de distribuição de aplicativos: APT
  - Disponibilidade de dados estatísticos: Popcon
  - Possibilidade de integração dos resultados
  - Popularidade

- Aplicativos GNU + kernel Linux
- Seleção e configuração de softwares
- Escolha do Debian GNU/Linux como plataforma
  - Esquema consistente de distribuição de aplicativos: APT
  - Disponibilidade de dados estatísticos: Popcon
  - Possibilidade de integração dos resultados
  - Popularidade



- Aplicativos GNU + kernel Linux
- Seleção e configuração de softwares
- Escolha do Debian GNU/Linux como plataforma
  - Esquema consistente de distribuição de aplicativos: APT
  - Disponibilidade de dados estatísticos: Popcon
  - Possibilidade de integração dos resultados
  - Popularidade



- Aplicativos GNU + kernel Linux
- Seleção e configuração de softwares
- Escolha do Debian GNU/Linux como plataforma
  - Esquema consistente de distribuição de aplicativos: APT
  - Disponibilidade de dados estatísticos: Popcon
  - Possibilidade de integração dos resultados
  - Popularidade



- Aplicativos GNU + kernel Linux
- Seleção e configuração de softwares
- Escolha do Debian GNU/Linux como plataforma
  - Esquema consistente de distribuição de aplicativos: APT
  - Disponibilidade de dados estatísticos: Popcon
  - Possibilidade de integração dos resultados
  - Popularidade

- Aplicativos GNU + kernel Linux
- Seleção e configuração de softwares
- Escolha do Debian GNU/Linux como plataforma
  - Esquema consistente de distribuição de aplicativos: APT
  - Disponibilidade de dados estatísticos: Popcon
  - Possibilidade de integração dos resultados
  - Popularidade



# Recomendação nas Distribuições [+]

- Cross-distribution Meeting on Application Installer
- Destaque para Debtags e Apt-xapian-index
- Fontes de informação no Debian
  - BTS, Popcon, PTS, UDD, Debtags
- Iniciativas descontinuadas
  - PopSuggest e Debommender

- Cross-distribution Meeting on Application Installer
- Destaque para Debtags e Apt-xapian-index
- Fontes de informação no Debian
  - BTS, Popcon, PTS, UDD, Debtags
- Iniciativas descontinuadas
  - PopSuggest e Debommender

- Cross-distribution Meeting on Application Installer
- Destaque para Debtags e Apt-xapian-index
- Fontes de informação no Debian
  - BTS, Popcon, PTS, UDD, Debtags
- Iniciativas descontinuadas
  - PopSuggest e Debommender

- Cross-distribution Meeting on Application Installer
- Destaque para Debtags e Apt-xapian-index
- Fontes de informação no Debian
  - BTS, Popcon, PTS, UDD, Debtags
- Iniciativas descontinuadas
  - PopSuggest e Debommender

- Cross-distribution Meeting on Application Installer
- Destaque para Debtags e Apt-xapian-index
- Fontes de informação no Debian
  - BTS, Popcon, PTS, UDD, Debtags
- Iniciativas descontinuadas
  - PopSuggest e Debommender

- Cross-distribution Meeting on Application Installer
- Destaque para Debtags e Apt-xapian-index
- Fontes de informação no Debian
  - BTS, Popcon, PTS, UDD, Debtags
- Iniciativas descontinuadas
  - PopSuggest e Debommender

### Desenvolvimentos do Trabalho

- Caracterização do problema
- Metodologia
- Plano de execução



### Desenvolvimentos do Trabalho

- Caracterização do problema
- Metodologia
- Plano de execução



### Desenvolvimentos do Trabalho

- Caracterização do problema
- Metodologia
- Plano de execução



- Modelagem de usuários e itens
- Ação do recomendador: encontrar itens relevantes
- Pontuação binária vs multi-valorada
  - Open Collaboration Services [+]
- Considerações sobre dependências
- Necessidades modeladas como funcionalidades



- Modelagem de usuários e itens
- Ação do recomendador: encontrar itens relevantes
- Pontuação binária vs multi-valorada
   Open Collaboration Services
- Considerações sobre dependências
- Necessidades modeladas como funcionalidades

- Modelagem de usuários e itens
- Ação do recomendador: encontrar itens relevantes
- Pontuação binária vs multi-valorada
  - Open Collaboration Services [+]
- Considerações sobre dependências
- Necessidades modeladas como funcionalidades

- Modelagem de usuários e itens
- Ação do recomendador: encontrar itens relevantes
- Pontuação binária vs multi-valorada
  - Open Collaboration Services [+]
- Considerações sobre dependência
- Necessidades modeladas como funcionalidades

- Modelagem de usuários e itens
- Ação do recomendador: encontrar itens relevantes
- Pontuação binária vs multi-valorada
  - Open Collaboration Services [+]
- Considerações sobre dependências
- Necessidades modeladas como funcionalidades



- Modelagem de usuários e itens
- Ação do recomendador: encontrar itens relevantes
- Pontuação binária vs multi-valorada
  - Open Collaboration Services [+]
- Considerações sobre dependências
- Necessidades modeladas como funcionalidades

- Coleta de dados
  - Popcon, UDD e Debtags
- Seleção de atributos
  - Dependência entre pacotes
  - Dados temporais
- Estratégias de recomendação
- Codificação
- Avaliação



- Coleta de dados
  - Popcon, UDD e Debtags
- Seleção de atributos
  - Dependência entre pacotes
  - Dados temporais
- Estratégias de recomendação
- Codificação
- Avaliação

- Coleta de dados
  - Popcon, UDD e Debtags
- Seleção de atributos
  - Dependência entre pacotes
  - Dados temporais
- Estratégias de recomendação
- Codificação
- Avaliação



#### Estratégias de recomendação

#### Baseada em conteúdo

- Técnicas: BM25 e Bayes ingênuo
- Dados: tags e descrições dos pacotes

#### Colaborativa

- Técnica: K-NN
- Dados: listas de pacotes instalados

#### Híbrida

- Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
- Meta-nível: colaboração por conteúdo
- Revezamento: decisão a partir da seleção de atributos
- Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Baseada em conteúdo
  - Técnicas: BM25 e Bayes ingênuo
  - Dados: tags e descrições dos pacotes
- Colaborativa
  - Técnica: K-NN
  - Dados: listas de pacotes instalados
- Híbrida
  - Cascata: refinamento de resultados a partir de dados demográficos e reputação dos itens
  - Meta-nível: colaboração por conteúdo
  - Revezamento: decisão a partir da seleção de atributos
  - Combinação dos resultados



- Linguagem Python
- Bibliotecas: Xapian e NLTK
- Documentação: Doxygen
- Controle de versão: Git

- Linguagem Python
- Bibliotecas: Xapian e NLTK
- Documentação: Doxygen
- Controle de versão: Git



- Linguagem Python
- Bibliotecas: Xapian e NLTK
- Documentação: Doxygen
- Controle de versão: Git

- Linguagem Python
- Bibliotecas: Xapian e NLTK
- Documentação: Doxygen
- Controle de versão: Git

- Testes preliminares: validação cruzada
- Seleção de estratégias e ajustes das soluções
- Consulta pública: survey eletrônico
  - Composição da identidade do usuário
  - Produção das recomendações por diversas estratégias
  - Apresentação dos resultados
  - Avaliação pelo usuário
  - 💿 Aplicação de métricas
- Compilação dos resultados globais

- Testes preliminares: validação cruzada
- Seleção de estratégias e ajustes das soluções
- Consulta pública: survey eletrônico
  - Composição da identidade do usuário
  - Produção das recomendações por diversas estratégias
  - Apresentação dos resultados
  - Avaliação pelo usuario
  - 📵 Aplicação de métricas
- Compilação dos resultados globais



- Testes preliminares: validação cruzada
- Seleção de estratégias e ajustes das soluções
- Consulta pública: survey eletrônico
  - Composição da identidade do usuário
  - Produção das recomendações por diversas estratégias
  - Apresentação dos resultados
  - Avaliação pelo usuário
  - Aplicação de métricas
- Compilação dos resultados globais



- Testes preliminares: validação cruzada
- Seleção de estratégias e ajustes das soluções
- Consulta pública: survey eletrônico
  - Composição da identidade do usuário
  - Produção das recomendações por diversas estratégias
  - Apresentação dos resultados
  - Avaliação pelo usuário
  - Aplicação de métricas
- Compilação dos resultados globais



- Testes preliminares: validação cruzada
- Seleção de estratégias e ajustes das soluções
- Consulta pública: survey eletrônico
  - Composição da identidade do usuário
  - Produção das recomendações por diversas estratégias
  - Apresentação dos resultados
    - Avaliação pelo usuário
  - Aplicação de métricas
- Compilação dos resultados globais



- Testes preliminares: validação cruzada
- Seleção de estratégias e ajustes das soluções
- Consulta pública: survey eletrônico
  - Composição da identidade do usuário
  - Produção das recomendações por diversas estratégias
  - Apresentação dos resultados
  - Avaliação pelo usuário
  - Aplicação de métricas
- Compilação dos resultados globais

- Testes preliminares: validação cruzada
- Seleção de estratégias e ajustes das soluções
- Consulta pública: survey eletrônico
  - Composição da identidade do usuário
  - Produção das recomendações por diversas estratégias
  - Apresentação dos resultados
  - Avaliação pelo usuário
  - Aplicação de métricas
- Compilação dos resultados globais



- Testes preliminares: validação cruzada
- Seleção de estratégias e ajustes das soluções
- Consulta pública: survey eletrônico
  - Composição da identidade do usuário
  - Produção das recomendações por diversas estratégias
  - Apresentação dos resultados
  - Avaliação pelo usuário
  - Aplicação de métricas
- Compilação dos resultados globais



- Testes preliminares: validação cruzada
- Seleção de estratégias e ajustes das soluções
- Consulta pública: survey eletrônico
  - Composição da identidade do usuário
  - Produção das recomendações por diversas estratégias
  - Apresentação dos resultados
  - Avaliação pelo usuário
  - Aplicação de métricas
- Compilação dos resultados globais



## Plano de Execução

Atividade	Jan	Fev	Mar	Abr	Mai	Jun	Jul
Preparação e realização da qualificação	Х	х					
Implementação de difer- entes estratégias e técni- cas	х	x	x	x			
Testes e ajustes na imple- mentação				Х	х		
Preparação e aplicação do <i>survey</i>	·				х	х	
Escrita da dissertação					Х	Х	Х



- Acompanhamento das discussões no FreeDesktop
- Desenvolvimento nos padrões OCS
- Conclusão pautada pelos resultados do survey
- Meta++: integração de um recomendador na estrutura do software-center



- Acompanhamento das discussões no FreeDesktop
- Desenvolvimento nos padrões OCS
- Conclusão pautada pelos resultados do survey
- Meta++: integração de um recomendador na estrutura do software-center



- Acompanhamento das discussões no FreeDesktop
- Desenvolvimento nos padrões OCS
- Conclusão pautada pelos resultados do survey
- Meta++: integração de um recomendador na estrutura do software-center



- Acompanhamento das discussões no FreeDesktop
- Desenvolvimento nos padrões OCS
- Conclusão pautada pelos resultados do survey
- Meta++: integração de um recomendador na estrutura do software-center



### Referências

- Araújo, T. C. Aplicação de Estratégias de Recomendação ao Domínio de Componentes de Software. 2011. Disponível em https://github.com/tassia/QualificacaoMestrado
- Araújo, T. C. AppRecommender: Application Recommender for GNU/Linux Systems. 2011. Disponível em https://github.com/tassia/AppRecommender