

LUCAS FERNANDES BRUNIALTI

**Biclusterização aplicada em Sistemas de
Recomendação baseados em Conteúdo Textual**

São Paulo

2014

LUCAS FERNANDES BRUNIALTI

Biclusterização aplicada em Sistemas de Recomendação baseados em Conteúdo Textual

Texto de Exame de Qualificação apresentado à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo como parte dos requisitos para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Orientador: Profa. Dra. Sarajane Marques Peres

Co-Orientador: Prof. Dr. Valdinei Freire da Silva

São Paulo

2014

Autorização para Reprodução

Ficha catalográfica

Folha de Aprovação

Texto de Exame de Qualificação de Mestrado sob o título “*Biclusterização aplicada em Sistemas de Recomendação baseados em Conteúdo Textual*”, apresentado por Lucas Fernandes Brunialti e aprovado em ___ de _____ de _____, em São Paulo, Estado de São Paulo, pela comissão examinadora constituída pelos doutores:

Prof. Dr. _____
Presidente

Instituição: _____

Prof. Dr. _____
Instituição: _____

Prof. Dr. _____
Instituição: _____

Resumo

BRUNIALTI, Lucas Fernandes. **Biclusterização aplicada em Sistemas de Recomendação baseados em Conteúdo Textual**. 2015. NúmeroDePáginas f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, Ano2.

Biclusterização representa uma estratégia de análise de dados com potencial para encontrar grupos de objetos similares considerando a correlação parcial existente entre os atributos descritivos dos mesmos. Esse potencial pode ser particularmente útil para Sistemas de Recomendação baseados em conteúdo, nos quais se faz necessária a sugestão de itens úteis, porém diversificados, para um usuário. Esta necessidade configura-se como um problema de serendipidade, uma das propriedades de Sistemas de Recomendação. O objetivo deste trabalho é analisar a aderência dos resultados provenientes de algoritmos de biclusterização, aplicados aos itens a serem recomendados, à meta de otimização da serendipidade. Para alcançar tal objetivo, este trabalho propõe um estudo da aplicação de algoritmos clássicos de biclusterização à conteúdo textual referente ao escopo de um Sistema de Recomendação de notícias. Ainda, este trabalho propõe a utilização de *essembles* como uma forma alternativa de encontrar biclusters. A hipótese defendida é que devido à análise de correlações parciais dos atributos descritivos dos dados, seria possível encontrar notícias com similaridades parciais entre si devido às particularidades presentes nas mesmas. Desta forma, uma mesma notícia que cobre mais de um contexto poderia ser recomendada a usuários que estariam, à princípio, interessandos em assuntos diferentes. Como forma de avaliação dos resultados obtidos, é proposta uma análise comparativa entre os resultados da recomendação obtida via biclusterização e os resultados de recomendação obtida via filtro colaborativo, onde o problema da serendipidade é reconhecidamente bem resolvido.

Palavras-chave: Biclusterização. Sistemas de Recomendação. Recomendação de Notícias. Mineração de Textos. *Essembles*.

Abstract

BRUNIALTI, Lucas Fernandes. **Work title**. 2015. NumberOfPages p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, Year2.

Write here the English version of your “Resumo ”...

Keywords: Keyword1. Keyword2. Keyword3. Etc.

Lista de Figuras

Lista de Tabelas

Sumário

1	Introdução	9
1.1	Apresentação do Problema de Recomendação	9
1.2	Apresentação do contexto de Biclusterização	10
1.3	Hipótese	11
1.4	Objetivos	11
1.5	Metodologia	12
1.6	Organização do documento	12
2	Conceitos Fundamentais	13
2.1	Sistemas de Recomendação	13
2.1.1	Tipos de Sistemas de Recomendação	14
2.1.1.1	Limitações e Problemas	15
2.1.2	Avaliação da Recomendação	16
2.2	Biclusterização	17
2.2.1	Tipo de biclustering	17
2.2.2	Algoritmos para biclustering	18
2.2.3	Avaliação de biclustering	18
2.2.4	Essembles para clusterização	19
2.3	Mineração de Textos	19
2.3.1	Etapas de pré-processamento	20
2.3.2	Representação de textos	21

3	Sistemas de Recomendação por Conteúdo e Aprendizado de Máquina	22
4	Proposta.....	24
4.1	Apresentação do corpus IG	24
4.1.1	Pré processamento do corpus IG	24
4.2	Estudos iniciais de biclustering no corpus IG	25
4.3	Próximos passos - cronograma.....	25
	Referências	27

Capítulo 1

Introdução

Sugestão: fazer uma introdução contextualizada em um problema de recomendação de notícias. Para esse problema, criar um cenário onde uma notícia poderia ser recomendada para dois usuários diferentes que estariam inicialmente navegando em notícias diferentes (de contextos diferentes). Mostrar que uma notícia pode estar relacionada a duas outras que por sua vez não estão relacionadas.

[illegible]

Texto texto texto texto texto texto texto texto texto texto texto texto
texto texto texto texto texto texto texto texto texto texto texto texto
texto texto texto texto texto texto texto texto texto texto texto texto
texto texto texto texto texto texto texto texto texto texto texto texto
texto texto texto texto texto texto texto texto texto texto texto texto
texto.

[illegible]

1.1 Apresentação do Problema de Recomendação

Apresentar de maneira mais formal (acima foi só um cenário), o problema de recomendação. Tentar caracterizar (aí acho que de maneira não formal necessariamente)

Capítulo 2

Conceitos Fundamentais

Este capítulo introduz os conceitos fundamentais para o entendimento dessa dissertação, fazendo um apanhado dos conceitos na área de Sistemas de Recomendação (seção 2.1), Biclusterização (seção ??) e Mineração de Texto (seção ??).

2.1 Sistemas de Recomendação

A grande maioria das pessoas que usam a Internet muito provavelmente já interagiram com algum Sistema de Recomendação (SR), por isso, o seu conceito é intuitivo. Porém, por ser uma área relativamente nova (handbook-1), muitos autores não usam uma definição que reflete a realidade dos SRs, isso acontece por ser uma área relativamente nova que teve um crescimento muito grande nos últimos anos (tese-chap2).

Resnick and Varian (1997), quem criou o termo Sistemas de Recomendação (Tese chap2, (Neumann, 2007)), argumentam que Sistemas de Recomendação servem para nos ajudar em processos de tomada de decisão do nosso dia-a-dia, como quais itens comprar, quais músicas ouvir, ou quais notícias ler. Além disso, Resnick and Varian (1997) provê uma taxonomia para definição de Sistemas de Recomendação:

- Conteúdo recomendado: os itens que são recomendados pelo Sistema de Recomendação, ex: produtos, músicas, notícias e/ou etc.
- Entrada dos usuários: as interações que os usuários realizam com os itens são a entrada para um SR, estas podem ser implícitas (ex: o usuário x leu a notícia y) ou explícitas (ex: o usuário x classificou o filme y como 5 estrelas).
- Target of recommendation: os itens recomendados podem ser diretamente para um usuário (personalizado), direcionados para um grupo de usuários ou todos os usuários (não-personalizado).
- Técnicas para recomendação (agregações): qual as estratégias e os algoritmos que os SRs usam para criar recomendações.

- Uso das recomendações: trata-se de como mostrar as recomendações para os usuários, ex: filtrando recomendações negativas, ordenando pelo fator numérico, etc.

Porém, definições mais recentes (Burke 2002 e 2007), descrevem SRs como qualquer sistema que produz recomendações personalizadas ou tem o efeito de guiar um usuário de modo personalizado, mostrando itens que possam ser interessantes para este usuário, dentro de uma grande quantidade de opções. Isso faz com que SRs que provêm recomendações não personalizadas deixem de adequar com a definição de SR. É provável que isso se deve ao fato que das estratégias usadas atualmente, que têm o foco de produzir recomendações personalizadas.

Formalmente, Burke 2002 e 2007 define SRs como I um conjunto de itens que podem ser recomendados e U um conjunto de usuários que as preferências são conhecidas, u um usuário para o qual as recomendações são geradas, e i algum item que queremos prever a preferência para u . Adomavicius estende a definição com uma função de utilidade f que mede o quão útil é o item i para o usuário u : $f : I \times U \rightarrow R$, em que R um conjunto ordenado de inteiros ou reais.

No entanto, existem tipos de SRs que não estimam f completamente, podendo otimizar funções auxiliares para gerar as recomendações para um usuário u (handbook-1).

Em síntese, um Sistema de Recomendação tem a função de auxiliar os usuários de uma aplicação a interagir com itens, provendo sugestões de quais itens interagir, baseando-se no histórico de interações desses usuários com esses itens.

2.1.1 Tipos de Sistemas de Recomendação

Para estimar f e chegar no conjunto ordenado R existem diversas estratégias, daí surgem os tipos de SRs. Os tipos de SRs diferem quanto ao domínio, dados usados para recomendação, algoritmos (handbook-1), e principalmente nas propriedades em que cada tipo se destaca.

Burke 2002 e 2007 provê uma taxonomia já considerada clássica (handbook-1), que categoriza os SRs em cinco diferentes tipos:

- *Filtragem colaborativo*, o primeiro tipo de SR que foi implementado (Resnick), tem como idéia básica encontrar outros usuários $u_{1..n}$ em U com preferências semelhantes à u , e então recomendar itens que $u_{1..n}$ interagiram e que u ainda não interagiu, estabelecendo alguma métrica para estimar f . Medidas geralmente usadas incluem

Correlação de Pearson e *Similaridade dos Cossenos*, também são usadas técnicas para redução de dimensionalidade, como *Decomposição de Valores Singulares* e *Fatorização de Matriz* (Livro).

- *Baseado em conteúdo*, é um dos tipos de SR que otimiza funções auxiliares à f . Descreve os itens por características possibilitando o uso de medidas de similaridade entre itens. Então, com as interações dos usuários de U sob itens em I , constrói um perfil de interesses para cada usuário. Neste caso são usadas técnicas de Recuperação de Informação (livro) para representar os itens e calcular similaridades entre itens, assim como técnicas de Aprendizado de Máquina Supervisionado e Não-supervisionado (livro, burke2002).
- *Baseado em Conhecimento*, tem o intuito de sugerir itens, de forma personalizada, baseando-se nas necessidades ou regras estabelecidas por um usuário u e nas características dos itens em I . São estabelecidas medidas de similaridade para estimar o quanto as necessidades do usuário match as recomendações (livro, burke07, handbook-1).
- *Híbrido*, é capaz de combinar as vantagens de cada tipo de SR descrito para suprir as limitações associadas à cada tipo. A dificuldade está em como combinar as diferentes técnicas de cada algoritmo (livro, burke07). Burke07 identificou 7 tipos de SRs Híbridos em uma revisão da literatura: *Pesagem*, atribui um peso para cada algoritmo; *Switching*, seleciona um dos algoritmos (ou tipos); *Mixed*, recomendações são mostradas em conjunto; *Combinação de características*, diferentes fontes são combinadas em apenas um algoritmo; *Feature Augmentation*, uma técnica é usada para computar características que servem de entrada para outra técnica; *Cascade*, é atribuído um grau de prioridade para cada algoritmo; *Meta-level*, uma técnica gera um modelo, que é usado como entrada para outras técnicas.

2.1.1.1 Limitações e Problemas

Cada um dos tipos de SRs descritos possuem vantagens e limitações independentes se comparados com os outros tipos ou não. Em livro, Adomavicius, Burke e handbook-3 são nomeadas os problemas de *novo usuário* (*user cold-start*), *novo item* (*item cold-start*), *esparsidade*, *sobre-especialização* e *análise limitada do conteúdo* referentes aos SRs.

Os problemas de *novo usuário* e *novo item* são similares, basicamente, enquanto o

Referências