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

Capítulo 1

Introdução

Sugestão: fazer uma introdução contextualizada em um probema 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.

Texto texto.

Texto texto.

Texto texto.

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)

a propriedade de serendipidade.

Texto texto.

1.2 Apresentação do contexto de Biclusterização

Apresentar o que é biclusterização, em linhas gerais e com uma pequena formulação. Falar que existem vários algoritmos.

Texto texto.

Texto texto.

Apresentar em linhas gerais a possibilidade de implementar bicluterização usando a alternativa de essemble.

Texto texto.

Texto texto texto texto texto texto texto texto texto texto texto texto texto texto

texto texto.

1.3 Hipótese

Um grande problema referente à SRs baseados em conteúdo é a sobre-especialização, que é a falta de recomendações com conteúdo novo ou inesperado à um usuário, ou a falta de serendipidade nas recomendações. Como a Biclusterização é capaz de expandir o espaço de busca, levando em consideração subconjuntos das características para a formação de clusters, a hipótese desse projeto é que a aplicação de algoritmos de biclusterização em SRs baseados em conteúdo pode amenizar o problema de sobre-especialização, melhorando a qualidade das recomendações e contribuindo para a área de SRs.

1.4 Objetivos

Colocar o objetivo de resolver o problema acima, otimizando a serendipidade, através do uso de biclusterização nos documentos (notícias).

Colocar objetivos especificos:

- levantamento do referencial teórico de sistemas de recomendação textual
  - levantamento do estado da arte (sua revisao sistemática)
- levantamento de referencial teórico de biclustering
- construção de um conjunto de dados
- estudo de processamento de texto (text mining)
- estudo das técnicas de biclustering aplicadas a texto
- estudo das técnicas de essemble aplicadas a text

1.5 Metodologia

Aqui explicar cada um dos passos. Explicar que o levantamento do referencial teórico através de análise exploratória. Explicar como é a revisão sistemática (um parágrafo). Explicar a motivação de tê-la realizado. Texto texto.

Explicar como o conjunto de dados foi construído. Explicar a base de cliques. Texto texto.

Explicar que as técnica de biclustering e essemble serão aplicadas aos textos processados e que uma avaliação será feita com base nas informações de cliques. Texto texto.

1.6 Organização do documento

Este documento é composto por X capítulos incluindo esta introdução. O Capítulo 1 ..... e apresentar cada um. Texto texto.

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 definicão 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 pra o qual as recomendações são geradas, e i algum item que queremos predizer a preferência para u. Adomavicius extende 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 à 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, informações usadas para recomendação, algoritmos (handobook-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 $u_{1,\ldots,n}$ em U, sendo $n < |U|$ e não necessariamente em ordem, com preferências semelhantes à u, e então recomendar itens que $u_{1,\ldots,n}$ 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 usados 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, é construído um perfil de interesses para cada usuário. As recomendações são feitas a partir da combinação do perfil de interesses de um usuário u com os itens em I que u ainda não interagiu. 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 esta 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 Augumentation, 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 Vantagens e desvantagens

Cada um dos tipos de SRs descritos possuem algumas limitações independentes se comparados com os outros tipos ou não. Em livro, Adomavicius, Burke e handbook-3 são nomeadas os problemas no desenvolvimento de SRs de *novo usuário* (*user cold-start*), *novo item* (*item cold-start*), *esparsidade*, *sobre-especialização* e *análise de conteúdo limitada* referentes aos SRs.

Os problemas de *novo usuário* e *novo item* são similares, basicamente, enquanto o primeiro se trata da dificuldade de gerar recomendações para novos usuários, o segundo se trata de gerar recomendações para novos itens. O problema de *novo usuário* esta presente como uma desvantagem nos SRs de filtragem colaborativa e baseado em conteúdo, pois o SR não conhece as preferências dos novos usuários, tendo dificuldade de construir um modelo que tem como base essas preferências. Já o problema de *novo item* é considerado uma vantagem para os SRs baseados em conteúdo, enquanto uma desvantagem para os SRs baseados em filtragem colaborativa. Como a filtragem colaborativa se baseia apenas nas interações de usuários em itens, um item novo, que não teve ou teve poucas interações, não será recomendado, diferentemente do SR baseado em conteúdo, que leva em consideração a representação do item para construir recomendações.

Contrariamente, o problema de *sobre-especialização* é uma vantagem para os SRs de filtragem colaborativa e uma desvantagem para os SRs baseados em conteúdo. A *sobre-especialização* diz respeito ao problema de sugerir apenas itens previsíveis para o usuário, por exemplo, se o usuário viu notícias apenas de esporte, ele já espera receber sugestões de notícias de esporte, porém este usuário muito provavelmente pode gostar de ler notícias de outras categorias. A capacidade do SR de sugerir notícias imprevisíveis é chamado de *serendipidade* (livro, handbook-3). SRs baseados em conteúdo sofrem desse problema pois combina o perfil de preferências de um usuário com os itens em I, restrigindo o espaço de busca para realizar a sugestão de itens. Enquanto isso, os SRs baseados em filtragem colaborativa são capazes de oferecer sugestões úteis e serendipitas, pois são capazes de ampliar o espaço de busca através da estratégia de sugerir itens que usuários semelhantes à u interagiram e que u ainda não interagiu.

Os SRs baseados em filtragem colaborativa são os únicos que sofrem do problema de *esparsidade*, que é o fato de usuários interagirem com apenas um pequeno subconjunto do conjunto de itens, tornando a matriz de interações ou preferências ($U \times I$) esparsa. Isso faz com que aumente a necessidade de ter uma grande quantidade de usuários, pois este tipo de SR necessita de intersecções nas interações de itens por usuários, para que seja possível encontrar usuários similares a um dado usuário.

Assim como os SRs de filtragem colaborativa, os SRs baseados em conteúdo sofrem de um problema único, que é a *análise de conteúdo limitada*. Este problema se refere à representação dos itens, que tem de ser suficiente para discriminá-los (handbook-3). Em SRs baseados em conteúdo é comum ter que limitar a representação de itens, tanto pelo número de características quanto pela modelagem. Por exemplo, no contexto

de notícias na web, se a representação for um vetor com o número de ocorrências de cada palavra, perdemos a relação entre as palavras, e também todo o conteúdo que não é texto, como imagens, vídeos, etc.

2.1.2 Avaliação da Recomendação

Diferentes técnicas foram mostradas e uma variedade de problemas relacionados à essas técnicas, mas como avaliar se uma estratégia adotada no desenvolvimento realmente esta sendo efetiva?

2.2 Biclusterização

Um parágrafo com definição geral. Texto texto.

2.2.1 Tipo de biclustering

Texto texto.

Texto texto.

Texto texto

texto texto.

Texto texto.

2.2.2 Algoritmos para biclustering

Texto texto.

Texto texto.

Texto texto.

2.2.3 Avaliação de biclustering

Texto texto.

Texto texto.

2.2.4 Essembles para clusterização

Texto texto.

Texto texto.

Texto texto.

2.3 Mineração de Textos

Uma definição geral. Texto texto.

Texto texto.

2.3.1 Etapas de pré-processamento

Texto texto.

Texto texto.

Texto texto.

Texto texto.

2.3.2 Representação de textos

Texto texto.

Texto texto.

Texto texto.

Texto texto.

Capítulo 3

Sistemas de Recomendação por Conte e Aprendizado de Máquina

Texto texto.

Texto texto.

Texto texto.

Texto texto.

Texto texto

texto texto.

Texto texto.

Texto texto.

Capítulo 4

Proposta

Texto texto.

4.1 Apresentação do corpus IG

O corpus do IG é composto por um conjunto de notícias $\{n_1, n_2, \ldots, n_m\}$ em que cada notícia n_i é representado pela tupla (t, st, b, c_i), onde t é o título, st o subtítulo, b o corpo da notícia e c_i um elemento do conjunto de canais $\{gente, ultimosegundo, delas, economia, esporte, saude, igay, deles, tecnologia, receitas, igirl, jovem, arena, luxo\}$, em que cada canal representa um assunto ou tópico de notícias.

O número total de notícias é $m = xxx$, ...

4.1.1 Pré processamento do corpus IG

Texto texto.

Texto texto

texto texto texto texto texto texto texto texto texto texto texto texto texto texto texto texto.

4.2 Estudos iniciais de biclustering no corpus IG

Texto texto.

Texto texto.

Texto texto.

Texto texto.

4.3 Próximos passos - cronograma

Texto texto

texto texto.

Texto texto.

Texto texto.

Referências