PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Diogo Silveira Mendonça

Análise Probabilística de Semântica Latente aplicada a sistemas de recomendação

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio.

Orientador: Prof. Ruy Luiz Milidiú

Rio de Janeiro

Setembro de 2008

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Diogo Silveira Mendonça

Análise Probabilística de Semântica Latente aplicada a sistemas de recomendação

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Ruy Luiz Milidiú
Orientador
Departamento de Informática – PUC-Rio

Prof. Daniel Schwabe
Departamento de Informática – PUC-Rio

Prof. Marcus Vinicius Soledade Poggi de Aragão
Departamento de Informática – PUC-Rio

Prof. José Eugenio Leal
Coordenador(a) Setorial do Centro Técnico Científico - PUC-Rio

Rio de Janeiro, 12 de Setembro de 2008

Diogo Silveira Mendonça

Recebeu seu título de Bacharel em Ciência da Computação pela Universidade Federal do Rio de Janeiro (UFRJ) em 2006. Sua experiência acadêmica inclui pesquisas na área de Inteligência Artificial e Engenharia de Software.

Ficha Catalográfica

Mendonça, Diogo Silveira

Análise Probabilística de Semântica Latente aplicada a Sistemas de Recomendação / Diogo Silveira Mendonça; orientador: Ruy Luiz Milidiú. – Rio de Janeiro: PUC, Departamento de Informática, 2008.

69 f. ; 30,0 cm

1. Dissertação (Mestrado em Informática) – Pontifícia Universidade Católica do Rio de Janeiro, 2008.

Inclui referências bibliográficas.

1. Informática – Teses. 2. Análise probabilística de semântica latente. 3. Aprendizado de máquina. 4. Sistemas de recomendação. 5. Recomendação de anúncios na web. 6. Recomendação de filmes. I. Milidiú, Ruy Luiz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

A todos que procuram consciência e sabedoria.

Agradecimentos

Gostaria de agradecer inicialmente a minha família, em especial a minha Mãe, minha Irmã e meu Pai, que tem sempre me apoiado e dado todo o suporte necessário para que eu possa estudar e me desenvolver. Sem eles dificilmente eu cursaria o mestrado.

Agradeço também a minha namorada Taliha, por me amar muito e por aturar e ouvir pacientemente todas as minhas reclamações durante o mestrado. Lindona te amo muito!

Agradeço ao meu orientador por esclarecer das dúvidas mais simples as mais complexas permitindo que eu aprofundasse o meu conhecimento neste mundo antes desconhecido chamado Aprendizado de Máquina.

Agradeço aos meus colegas de mestrado, que o tornaram mais divertido e me ajudaram nos momentos de dificuldade. São eles: Andrew, Rodinei, Túlio, Sérgio, Márcio, Bruno, Ricardo, Eduardo, Roberto, Majowka e Cícero.

Agradeço também ao CNPq e ao Governo Federal, que apesar de todos os seus problemas e defeitos investe em pesquisa, o que possibilitou a realização do meu mestrado.

Agradeço finalmente a todos que estiveram próximo de mim e me apoiaram de forma direta ou indireta, mesmo que não estejam conscientes disto.

A todos Muitíssimo Obrigado!

Resumo

Mendonça, Diogo Silveira. **Análise Probabilística de Semântica Latente aplicada a sistemas de recomendação.** Rio de Janeiro, 2008. 69p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Os sistemas de recomendação são um tema de pesquisa constante devido a sua grande quantidade de aplicações práticas. Estes sistemas podem ser abordados de diversas maneiras, sendo uma das mais utilizadas a filtragem colaborativa, em que para recomendar um item a um usuário são utilizados dados de comportamento de outros usuários. Porém, nem sempre os algoritmos de filtragem colaborativa atingem níveis de precisão necessários para serem utilizados em aplicações reais. Desta forma este trabalho tem como objetivo avaliar o desempenho da análise probabilística de semântica latente (PLSA) aplicado a sistemas de recomendação. Este modelo identifica grupos de usuários com comportamento semelhante através de atributos latentes, permitindo que o comportamento dos grupos seja utilizado na recomendação. Para verificar a eficácia do método, apresentamos experimentos com o PLSA utilizando os problemas de recomendação de anúncios na web e a recomendação de filmes. Evidenciamos uma melhoria de 18,7% na precisão da recomendação de anúncios na web e 3,7% de melhoria no erro quadrático sobre a Média das Médias para o corpus do Netflix. Além dos experimentos, o algoritmo foi implementado de forma flexível e reutilizável, permitindo adaptação a outros problemas com esforço reduzido. Tal implementação também foi incorporada como um módulo do LearnAds, um framework de recomendação de anúncios na web.

Palavras-chave

Análise probabilística de semântica latente, aprendizado de máquina, sistemas de recomendação, recomendação de anúncios na web, recomendação de filmes.

Abstract

Mendonça, Diogo Silveira. **Probabilistic Latent Semantic Analysis applied to recommender systems.** Rio de Janeiro, 2008. 69p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Recommender systems are a constant research topic because of their large number of practical applications. There are many approaches to address these problems, one of the most widely used being *collaborative filtering*, in which in order to recommend an item to a user, data of other users' behaviors are employed. However, collaborative filtering algorithms do not always reach levels of precision required for the use in real applications. Within this context, the present work aims to evaluate the performance of the probabilistic latent semantic analysis (PLSA) applied to recommender systems. This model identifies groups of users with similar behaviors through latent attributes, allowing the use of these behaviors in the recommendation. To check the effectiveness of the method, there were presented experiments with problems of both web ad recommending and film recommending. An improvement of 18,7% were found in the accuracy of the recommendation of ads on the web and we also found 3.7% of improvement in Root Mean Square Error over the Means of Means baseline system for the Netflix corpus. Apart from the aforementioned experiments, the algorithm was implemented in a flexible and reusable way, allowing its adaptation to other problems with reduced effort. This implementation has also been incorporated as a module of LearnAds, a framework for the recommendation of ads on the web.

Keywords

Probabilistic latent semantic analysis, machine learning, recommender systems, web advertisement recommendation, films recommendation.

Sumário