



Miguel Angelo Gaspar Pinto

**Sistema Híbrido de Recomendação de Produtos com
Uso de Filtros Colaborativos e Números Fuzzy**

Tese de Doutorado

Tese apresentada ao programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica da PUC-Rio como parte dos requisitos parciais para obtenção do título de Doutor em Engenharia Elétrica.

Orientador: Prof. Ricardo Tanscheit

Co-orientadora: Profa. Marley Maria Bernardes Rebuzzi Vellasco

Rio de Janeiro
Setembro de 2011



Miguel Angelo Gaspar Pinto

Sistema Híbrido de Recomendação de Produtos com uso de Filtros Colaborativos e Números Fuzzy

Tese apresentada como requisito parcial para obtenção do grau de Doutor pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Ricardo Tanscheit
Orientador

Departamento de Engenharia Elétrica – PUC-Rio

Prof. Marley Maria Bernardes Rebuzzi Vellasco
Co-Orientadora

Departamento de Engenharia Elétrica – PUC-Rio

Prof. Karla Tereza Figueiredo Leite
UEZO

Prof. Jorge Luís Machado do Amaral
UERJ

Prof. Christian Nunes Aranha
Cortex

Prof. Juan Guillermo Lazo Lazo
ICA/DEE/PUC-Rio

Prof. Roxana Jimenez Contreras
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Emmanuel Piseces Lopes Passos
Graall Consultoria Empresarial Ltda

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico

Rio de Janeiro, 27 de setembro de 2011

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Miguel Angelo Gaspar Pinto

Graduou-se em Engenharia de Controle e Automação (Pontifícia Universidade Católica) em 2004. Trabalhou por meio ano na área de controle antes de começar mestrado na área de Processamento de Sinais e Controle na Pós-Graduação da PUC-Rio. Participou de congressos na área de controle e inteligência artificial. Atualmente trabalha em projetos de inovação em empresa própria. Suas áreas de interesse abrangem robótica, controle de sistemas, visão computacional e inteligência artificial.

Ficha Catalográfica

Pinto, Miguel Angelo Gaspar

Sistema híbrido de recomendação de produtos com uso de filtros colaborativos e números fuzzy / Miguel Angelo Gaspar Pinto; orientador: Ricardo Tanscheit; co-orientadora: Marley Maria Bernardes Rebuzzi Vellasco 2011.

129 f. ; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2011.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Marketing. 3. Números fuzzy. 4. Filtros colaborativos. 5. Recomendação. 6. Algoritmos baseados em conteúdo. I. Tanscheit, Ricardo. II. Vellasco, Marley Maria Bernardes Rebuzzi. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Dedico esta tese ao meu pai, que não me deixou desistir de terminá-la apesar das pressões para o contrário. Dedico também a minha mãe que faleceu e que espero esteja orgulhosa por mais este feito.

Agradecimentos

Agradeço aos meus orientadores pela enorme paciência que tiveram. Agradeço também a minha esposa pela compreensão e por aceitar os vários fins de semana que deixei de aproveitar com ela para completar este feito.

Resumo

Pinto, Miguel Ângelo Gaspar; Tanscheit, Ricardo (Orientador); Vellasco, Marley Maria Bernardes Rebuzzi (Co-orientadora). **Sistema Híbrido de Recomendação de Produtos com uso de Filtros Colaborativos e Números Fuzzy**. Rio de Janeiro, 2011. 129p. Tese de Doutorado. Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

O varejo virtual tem sido um importante setor para dinamização da economia, cujo valor das transações em 2010 ficou em torno de R\$10,6 bilhões. As lojas nesse segmento não possuem restrição de clientes ou de estoque, porém possuem consumidores pouco pacientes com várias outras lojas a sua disposição, sendo necessário que o item de seu interesse seja encontrado visível rapidamente. Buscando resolver este problema, foram desenvolvidos algoritmos de recomendação capazes de gerar listagens de produtos que fossem direcionados ao usuário. Os algoritmos de filtragem colaborativa são amplamente usados no varejo virtual, porém eles apresentam problemas devido a escala e esparsidade do banco de dados. Algoritmos baseados em conteúdo podem apresentar menor sensibilidade ao tamanho da base de dados, porém sua efetividade depende da existência de dados de usuários que comumente não estão presentes. Nesta tese, propõe-se um algoritmo híbrido que utiliza tanto a filtragem colaborativa quanto um algoritmo baseado em conteúdo para permitir boas recomendações em bases de dados esparsas e de grande porte. O algoritmo baseado em conteúdo faz uso de números fuzzy e técnicas de marketing para guiar sua recomendação apenas com base nos itens comprados pelo usuário, sem necessidade de quaisquer outros dados pessoais do usuário. O algoritmo proposto foi testado em bases de dados sintética e real, sendo comparado com um filtro colaborativo padrão para avaliar seu desempenho. Os resultados obtidos demonstram que o algoritmo híbrido proposto apresentou um desempenho superior ao do filtro colaborativo padrão em ambas as base de dados, apresentando invariância à esparsidade da base de dados.

Palavras-chave

Marketing; Números Fuzzy; Filtros Colaborativos; Recomendação, Algoritmos baseados em conteúdo.

Abstract

Pinto, Miguel Ângelo Gaspar; Tanscheit, Ricardo (Advisor); Vellasco, Marley Maria Bernardes Rebuzzi (Co-advisor). **Hybrid Recommendation System Based on Collaborative Filtering and Fuzzy Numbers**. Rio de Janeiro, 2011. 129p. PhD Thesis – Departamento de Engenharia Elétrica, Pontificia Universidade Católica do Rio de Janeiro.

The virtual retail has been an important sector at Brazilian economy, being a \$6.23 billion market in 2010, having 30% expansion on that period. The companies in such segment don't have client or product restrictions due to physical limitations. On the other hand, the consumers of this kind of retail have several options to buy and little patience to keep searching on the same website. The companies need to define which item will be shown to the consumer before he leaves for the next competitor. Several recommendation algorithms were developed to generate products list directed to the consumer. Nowadays the algorithms for collaborative filtering are well spread in virtual retail, but they have problems caused exactly by the huge quantity of data that exist on virtual retail. Content based algorithms are less sensitive to the size of the database, but their effectiveness depends on the existence of user data, which usually are not available. This thesis proposes a hybrid algorithm which uses both collaborative filtering and a content based algorithm to allow recommendations in huge sparse databases. The content base algorithm uses fuzzy numbers and marketing techniques to guide the recommendation using only the items brought by the user, without the need for further personal data from the consumer. The proposed algorithm was tested in both artificial and real databases, compared with a benchmark collaborative filter. The collected results show that the proposed hybrid algorithm provides superior performance than the benchmark collaborative filter in both databases, generating good results and presenting sparsity invariance. The proposed algorithm also solves problems of initialization, neighborhood transitivity and in cases when new users or items are inserted on database.

Keywords

Marketing; Fuzzy Numbers; Collaborative Filters; Recommendation, Content Based Algorithms

Sumário

1 INTRODUÇÃO	14
1.1. Motivação	14
1.2. Objetivos do Trabalho	16
1.3. Contribuições da Tese	17
1.4. Estrutura da Tese	17
2 MARKETING	19
2.1. Introdução	19
2.2. Definição	19
2.3. Segmentação, Seleção e Posicionamento	21
2.3.1. Segmentação	21
2.3.2. Seleção	23
2.3.3. Posicionamento	23
2.3.4. Marketing e Métodos de Apoio à Decisão	24
2.4. Comportamento do Consumidor	25
2.4.1. Fontes de Influências	25
2.4.2. Orçamento Familiar	27
2.4.3. Heurísticas	29
2.4.4. CRM e Marketing de Relacionamento	29
3 FILTRAGEM COLABORATIVA	31
3.1. Introdução	31
3.2. Recomendações Baseados em Regras de Associação	31
3.3. Filtros Colaborativos	32
3.3.1. Algoritmos Baseados em Memória	35
3.3.2. Algoritmos Baseados em Modelo	44
3.3.3. Técnicas de Filtragem Colaborativa Híbridas	49
3.4. Desafios de Filtros Colaborativos	51
3.4.1. Esparsidade dos Dados	52
3.4.2. Escalabilidade dos Dados	53
3.4.3. Sinônimos	54
3.4.4. Ovelha Cinza	54
3.4.5. Má Fé	55
3.4.6. Outros Desafios	55
3.5. Métricas de Avaliação	55
3.6. Bancos de Dados para Filtragem Colaborativa	61
3.7. Resumo Filtros Colaborativos	62
4 RECOMENDADORES BASEADOS EM CONTEÚDO	64
4.1. Introdução	64
4.2. Recomendações Baseados em Conteúdo	65
4.3. Recomendador Baseado em Conteúdo por Números Fuzzy	66
4.3.1. Modelagem	67
4.3.2. Similaridade entre números Fuzzy	71
4.3.3. Arquitetura do Sistema Hosseinpour	73
4.3.4. Vantagens e Desvantagens	75

5 FILTRO HIBRIDO FUZZY	77
5.1. Introdução	77
5.2. Recomendador baseado em conteúdo	78
5.2.1. Posicionamentos Compostos	79
5.2.2. Posicionamentos de Marcas	80
5.2.3. Posicionamentos de Usuários	81
5.2.4. Definindo a escolha do recomendador	82
5.3. Algoritmo de Filtragem Colaborativa	83
5.4. Hibridização	84
5.5. Aperfeiçoando o Método Fuzzy	87
5.6. Redução do Espaço de Busca	90
5.7. Avaliando o Modelo	91
5.8. Esquema das Bases de Dados	93
5.9. Definição das Bases de Dados	95
5.9.1. Regras de Criação	96
6 RESULTADOS EXPERIMENTAIS	104
6.1. Introdução	104
6.2. Treinamento na Base de Dados Sintética	105
6.2.1. FC Item-Item (Colaborativo Item-para-Item)	106
6.2.2. Filtro Fuzzy (Algoritmo Fuzzy Individual)	108
6.2.3. FC Categórico (Filtro Colaborativo de Categorias)	110
6.2.4. FILTRO HIBRIDO FUZZY	112
6.2.5. Comparações entre os Algoritmos	114
6.3. Resultados com a Base de Dados Real	116
7 Conclusões e Trabalhos Futuros	120
7.1. Conclusões	120
8 Referências bibliográficas	123

Lista de figuras

Figura 1 - Segmentação, Seleção e Posicionamento	21
Figura 2 - Segmentação de Mercado	22
Figura 3 – Posicionamento e os 4P's.....	24
Figura 4 - Processo de segmentação e posicionamento gera produtos com características diferentes.....	25
Figura 5 – Matriz de Pontuação (A) com exemplo do modelo vetor-espaco em FC baseados em usuários.....	36
Figura 6 – Isolando itens co-avaliados e cálculo de similaridade.	37
Figura 7 – Algoritmo de filtragem colaborativa baseada em item. Processo de geração de predição ilustrado por 5 vizinhos.	38
Figura 8 – Similaridade para FC baseados em itens ($w_{i,j}$) calculo baseado nos itens co-avaliados i e j pelos usuários 2, l e n.....	41
Figura 9 – Curvas de distribuição dos elementos relevantes (b) e não relevantes (a). O valor de limiar (t) determina a taxa de precisão (acerto) eo o ruído (erro) do sistema. [Sampaio 2006]	59
Figura 10 - Exemplo de uma curva ROC.	60
Figura 11 – Número fuzzy triangular	67
Figura 12 - Função de Pertinência para números fuzzy triangulares	68
Figura 13 – Interface com o usuário.....	73
Figura 14 – Sistema de recomendação por números fuzzy.	74
Figura 15 – Processo híbrido de filtragem em duas etapas.	86
Figura 16 – Filmes e Posicionamentos	92
Figura 17 – Esquema básico do banco de dados sintética.....	95
Figura 18- FC Item-Item variação com o valor de Top-N para diversos valores de esparsidade do banco de dados (número de usuários).....	107
Figura 19- FC Item-Item variação com o valor de esparsidade (número de usuários) comparada em diversos valores de Top-N	107
Figura 20- Filtro Fuzzy variação com o valor de Top-N para diversos valores de esparsidade do banco de dados (número de usuários).....	108
Figura 21- Filtro Fuzzy variação com o valor de esparsidade (número de usuários) comparada em diversos valores de Top-N.....	109
Figura 22- FC Categórico em variação com o valor de Top-N para diversos valores de esparsidade do banco de dados (número de usuários).....	110

Figura 23- FC Categórico em variação com o valor de esparsidade (número de usuários) comparada em diversos valores de Top-N	111
Figura 24- Filtro Híbrido Fuzzy em variação com o valor de Top-N para diversos valores de esparsidade do banco de dados (número de usuários).	112
Figura 25- Filtro Híbrido Fuzzy em variação com o valor de esparsidade (número de usuários) comparada em diversos valores de Top-N	113
Figura 26 – Comparação entre recomendadores para Top-6.....	115
Figura 27 – Comparação entre recomendadores para 20.000 usuários	115
Figura 28- Comparação entre Filtro Fuzzy e FC Item-Item – Precisão, Revocação e F1 na base de dados MovieLens	118

Lista de tabelas

Tabela 1 - Exemplo de distribuição orçamentária de pesquisa feita pelo IBGE - Distribuição das despesas monetária e não-monetária média mensal familiar, por anos de estudo da referência da família.....	28
Tabela 2 – Exemplo de matriz de avaliação	40
Tabela 3 – Categorização de itens para revocação e precisão	57
Tabela 4 – Resumo de Técnicas de Filtragem Colaborativa com suas vantagens e desvantagens.	63
Tabela 5 – Termos Lingüísticos dos números fuzzy triangulares	69
Tabela 6 – Tabela de comparação de utilidade de celulares para mulheres (pesquisa Mattiota)	71
Tabela 7 – Classes sociais no Brasil – ABEP 2010.....	97

Lista de quadros

Quadro 1 – Definição de posicionamento do usuário baseado em itens comprados.....	88
Quadro 2 - Algoritmo de Regra de Compra	103