Pontifícia Universidade Católica
do Rio de Janeiro

Demetrius Costa Rapello

Sistema de recomendação de segundo nível para suporte à produção de matérias jornalísticas

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio.

Orientador: Prof. Marco Antonio Casanova

Rio de Janeiro
Março de 2012

Pontifícia Universidade Católica
do Rio de Janeiro

Demetrius Costa Rapello

Sistema de recomendação de segundo nível para suporte à produção de matérias jornalísticas

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Marco Antonio Casanova
Orientador
Departamento de informática - PUC-Rio

Prof. Antonio Luz Furtado
Departamento de informática - PUC-Rio

Prof. Ruy Luiz Milidiú
Departamento de informática - PUC-Rio

Prof. Karin K. Breitman
Departamento de informática - PUC-Rio

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico – PUC-Rio

Rio de Janeiro, 15 de março de 2012

Demetrius Costa Rapello

Graduou-se em Ciência da Computação pela Universidade de Augusto Motta (UNISUAM) em dezembro de 2000. Tem experiência na área de Ciência da Computação, com ênfase em Desenvolvimento de Software. Tem trabalhado em analise de sistemas desde 1997.

Ficha Catalográfica

Demetrius Costa Rapello

Sistema de recomendação de segundo nível para suporte à produção de matérias jornalísticas / Demetrius Costa Rapello; orientador: Marco Antonio Casanova. - Rio de Janeiro: PUC-Rio, Departamento de Informática, 2012.

v., 72 f.: il. ; 29,7 cm

Dissertação de Mestrado - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática

Referências bibliográficas incluídas.

Sistemas de Recomendação; Recuperação da Informação; Extração de Entidades Nomeadas;

CDD: 004

A Deus, aos meus pais, a minha esposa e aos meus filhos.

Agradecimentos

A Deus pelo apoio incondicional, conforto de coração e paz de espírito que tanto foram importantes para alcançar este objetivo.

Aos meus pais, Tarciso Rapello e Maria da Conceição Costa Rapello, pelo apoio, educação, carinho e dedicação.

A minha esposa Gabriela Barbosa da Silva Rapello, pelo apoio constante, carinho e compreensão.

Aos meus filhos Gabriel e Matheus que entenderam a ausência do pai com carinho e compreensão.

Ao meu orientador, prof. Marco Antonio Casanova, por sua dedicação, motivação, ensinamentos e orientação.

À Globo.com, pelo financiamento e auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Aos professores da Comissão examinadora.

A todos os amigos e familiares que de alguma forma contribuíram para a realização deste trabalho.

Resumo

Rapello, Demetrius Costa; Casanova, Marco Antonio. **Sistema de recomendação de segundo nível para suporte à produção de matérias jornalísticas.** Rio de Janeiro, 2011. 72p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Sistemas de recomendação têm sido amplamente utilizados pelos grandes portais na Web, em decorrência do aumento do volume de dados disponíveis na Web. Tais sistemas são basicamente utilizados para sugerir informações relevantes para os seus usuários. Esta dissertação apresenta um sistema de recomendação de segundo nível para auxiliar equipes de jornalistas de portais de notícias no processo de recomendação de notícias relacionadas para os usuários do portal. O sistema é chamado de *segundo nível* pois apresenta recomendações aos jornalistas para que, por sua vez, geram recomendações aos usuários do portal. O modelo seguido pelo sistema consiste na recomendação de notícias relacionadas com base em características extraídas do próprio texto da notícia original. As características extraídas permitem a criação de consultas contra um banco de dados de notícias anteriormente publicadas. O resultado de uma consulta é uma lista de notícias candidatas à recomendação, ordenada pela similaridade com a notícia original e pela data de publicação, que o editor da notícia original manualmente processa para gerar a lista final de notícias relacionadas.

Palavras-chave

Sistemas de Recomendação; Recuperação da Informação; Extração de Entidades Nomeadas.

Abstract

Rapello, Demetrius Costa; Casanova, Marco Antonio. **Second level recommendation system to support news editing.** Rio de Janeiro, 2011. 72p. MSc Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Recommendation systems are widely used by major Web portals due to the increase in the volume of data available on the Web. Such systems are basically used to suggest information relevant to their users. This dissertation presents a second-level recommendation system, which aims at assisting the team of journalists of a news Web portal in the process of recommending related news for the users of the Web portal. The system is called *second level* since it creates recommendations to the journalists Who, in turn, generate recommendations to the users. The system follows a model based on features extracted from the text itself. The extracted features permit creating queries against a news database. The query result is a list of candidate news, sorted by score and date of publication, which the news editor manually processes to generate the final list of related news.

Keywords

Recommendation Systems; Information Retrieval; Named Entity Extraction.

Sumário

1. Introdução 13

1.1. Motivação 13

1.2. Objetivo da dissertação 14

1.3. Organização da dissertação 14

2. Conceitos, técnicas e trabalhos relacionados 16

2.1. Conceitos e Técnicas 16

2.1.1. Vector Space Model 16

2.1.2. PageRank 17

2.1.3. Classificador bayesiano ingênuo 17

2.1.4. Part-of-speech Tagging 19

2.2. Trabalhos relacionados 19

2.2.1. Sistemas de recomendação 19

2.2.2. Ferramentas para extração de entidades nomeadas 22

2.2.3. Outras Ferramentas 24

3. GRNews – Sistema de recomendação de segundo nível para suporte a produção de matérias jornalísticas 29

3.1. Visão geral do sistema GRNews 29

3.2. Extrator de *features* 32

3.2.1. Discussão geral sobre *features* 32

3.2.2. Termos mais frequentes 33

3.2.3. Tags HTML informativas 34

3.2.4. Texto em títulos de vídeos e fotos 35

3.2.5. Reconhecimento de entidades nomeadas 35

3.3. Seleção de candidatos 44

3.3.1. Definição dos critérios de filtragem e ordenação 44

3.3.2. Combinação de *features* 46

3.3.3. Recuperação das candidatas 47

3.4. Recomendação 49

3.4.1. Fator de Similaridade 49
3.4.2. Fator de popularidade 50
3.5. Comentários do projeto 50

4. Experimento 53
4.1. O corpus 53
4.2. Critério para avaliação do sistema 54
4.3. Resultados obtidos 55

5. Conclusões e trabalhos futuros 61
5.1. Resumo do trabalho 61
5.2. Principais contribuições 62
5.3. Limitações do sistema 62
5.4. Trabalhos futuros 63

6. Bibliografia 64
7. APÊNDICE A 68

Lista de imagens

Figura 1 – Modelo do classificador.....18
Figura 2 – GRNews arquitetura.....29
Figura 3 – Interface Web do formulário de matéria.....30
Figura 4 – Componente de matérias recomendadas.....32
Figura 5 – Exemplo de matéria.....35
Figura 6 – Processo de extração de entidades.....37
Figura 7 – Acurácia do *POS-tagger*.....40
Figura 8 – Resultado do classificador 1.....41
Figure 9 – Resultado do classificador 2.....41
Figura 10 – Exemplo de matéria publicada.....47
Figure 11 – Fórmula de Precisão.....55
Figura 12 – Tabela de acurácia por editoria com 5 recomendações e sem o fator de similaridade.....69
Figura 13 – Tabela de acurácia por editoria com 5 recomendações com o fator de similaridade.....70
Figura 14 – Tabela de acurácia por editoria com 10 recomendações e sem o fator de similaridade.....71
Figura 15 – Tabela de acurácia por editoria com 10 recomendações e com o fator de similaridade.....72

Lista de tabelas

Tabela 1 – Comparativo de extratores de entidades 23
Tabela 2 – Exemplo de frases substantivas.....27
Tabela 3 – Campos do formulário de matéria 31
Tabela 4 – Formatos de *n-grams* 33
Tabela 5 – Lista de *n-grams*.....38
Tabela 6 – Corpora para reconhecer entidades.....39
Tabela 7 – Comparativo de entidades extraídas.....42
Tabela 8 – Comparativo entre extratores.....43
Tabela 9 – Exemplo de combinação de features 46
Tabela 10 – Representação do corpus 54
Tabela 11 – Acurácia por *feature* com 5 recomendações sem similaridade 56
Tabela 12 – Acurácia por *feature* com 5 recomendações com similaridade 58
Tabela 13 – Acurácia por feature com 10 recomendações sem similaridade 59
Tabela 14 – Acurácia por *feature* com 10 recomendações com similaridade 60

Lista de funções

Função 1 – Fórmula de distância dos cossenos 17
Função 2 – Exemplo de pagerank 17
Função 3 – Função de recomendação do PURE 25
Função 4 – Cálculo do Z-score no PURE 26
Função 5 – Cálculo de score do Lucene 45