PUC

INF3006 – Exame de proposta de dissertação

Sistema de recomendação para suporte a produção de artigos do portal techtudo

Demetrius Costa Rapello



Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

RUA MARQUÊS DE SÃO VICENTE, 225 - CEP 22453-900

RIO DE JANEIRO - BRASIL

Proposta de dissertação

Sistema de recomendação para suporte a produção de artigos do portal techtudo

Demetrius Costa Rapello

demetrius.rapello@gmail.com

Proposta de Dissertação apresentada ao Programa de Pós-Graduação em Informática da PUC-Rio como requisito parcial para obtenção do grau de Mestre em Informática

Orientador: Marco Antônio Casanova

Sumário

1 Introdução 1

2 Enunciado do Problema 2

3 Trabalhos Relacionados 3

4 Estado da Arte 6

5 Proposta 8

5.1 Sistema de Publicação de Matérias 8

5.2 Sistema especialista 8

5.3 Verificação 8

5.3.1 Avaliação da precisão do algoritmo 9

5.4 Informação sobre o corpus utilizado 9

6 Esboço do documento final 11

7 Plano de ação (esboço da solução) 12

8 Cronograma 17

1. Introdução

Podemos observar na última década um forte crescimento da internet, seja pelo número de usuários que corresponde a cerca de 30% da população mundial [14], seja pelo volume de páginas indexadas nos principais buscadores, que aproximasse da casa dos 20 bilhões [13]. Com estas informações, podemos supor que a oferta de conteúdo disponível na internet é muito superior a capacidade de consumo dos usuários.

Com o volume de conteúdo crescendo, os grandes portais de internet tem o desafio de oferecer mecanismos que auxiliem seus usuários no consumo de conteúdo.

Sistemas de recomendação tem sido largamente utilizados para atender a esta necessidade, pois em geral, estes sistemas são construídos com o intuito de predizer as necessidades do usuário de modo que se possa oferecer conteúdos com uma maior probabilidade de aproveitamento.

Para aprender a predizer as necessidades dos usuários, os sistemas de recomendação necessitam de acesso ao feedback dos usuários quando estes estão consumindo um determinado conteúdo. Este feedback pode ser dado de forma explicita quando um usuário informa ao sistema quais são as suas preferencias, ou de forma implícita quando o sistema descobre as preferencias do usuário baseado em suas interações.

Diante deste cenário, propomos um sistema de recomendação que aprenda as necessidades dos usuários baseado no fluxo de navegação entre as matérias para sugerir conteúdos relacionados aos novos artigos produzidos.

1. Enunciado do Problema

Com a necessidade de recomendação de conteúdo relacionado, os editores do portal techtudo passaram a ter uma segunda função. Atualmente, além da criação da nota ou artigo sobre um determinado tema, eles ainda precisam relacionar outras notas e ou artigos no sistema de recomendação existente. Deste modo o processo de produção de conteúdo se torna lento e a totalidade do tempo do editor, que deveria ser para confecção mais elaborada da matéria, passar a ser destinado a um trabalho manual de mineração de texto para realização de relacionamentos de conteúdos.

Para gerar recomendações de conteúdos para uma matéria, o editor precisa utilizar um fluxo de trabalho que passa pelas seguintes etapas:

1 - Identificação de palavras/frases chaves relevantes ao tema.

2 - Consulta de conteúdos anteriores na base de conhecimento utilizando as chaves.

3 - Reconhecimento da similaridade entre os conteúdos e o tema.

4 - Relacionamento manual dos conteúdos.

1. Trabalhos Relacionados

O projeto PURE [12] é um sistema para recomendação de artigos médicos que utiliza o princípio de recomendação baseado na filtragem de conteúdo. Em linhas gerais, o sistema executa uma classificação dos artigos preferidos do usuário para então recomendar outros artigos para leitura.

O sistema PURE pode ser entendido pelo fluxo de operações a seguir:

1. O usuário acessa o sistema para informar os artigos do seu interesse. Estes artigos são armazenados na base de dados do PURE.
2. Um sistema de aprendizado de máquina é aplicado para extrair as preferências do usuário com base nos seus artigos de interesse.
3. O sistema PURE consulta a base de dados PubMed para baixar os novos artigos publicados.
4. Os artigos baixados da PubMed são ordenados com base no modelo treinado com as preferências do usuário.
5. Os artigos são então apresentados para o usuário.

O propósito do PURE desenvolvido por TAKASHI YONEYA e HIROSHI MAMITSUKA, é oferecer uma interface que facilite o processo de busca de artigos relevantes pela comunidade científica na base de dados PubMed. A base de dados PubMed mantém um grande acervo de artigos de biologia e medicina com um volume diário de atualização da ordem de centenas artigos. Diante deste cenário, os autores perceberam que um sistema de recomendação baseado em filtragem de conteúdo poderia auxiliar a comunidade científica no trabalho de consulta de novos artigos.

Os seguintes módulos fazem parte do sistema PURE:

**Interface para registro de artigos de interesse do usuário**

Para utilizar o sistema, o usuário precisa registrar os seus artigos de interesse na base de dados do PURE. Para esta atividade o usuário acessa a interface web do sistema e seleciona os artigos de sua preferência em uma listagem.

Os artigos selecionados são então gravados no perfil do usuário e armazenados no banco de dados do PURE.

O usuário tem a permissão de adicionar novos arquivos e alterar sua lista de interesse.

**Treinamento do modelo probabilístico baseado nos artigos de interesse do usuário**

Os artigos de interesse do usuário são utilizados para a concepção de um modelo probabilístico que procura identificar as preferências do usuário para novos artigos. Esse modelo é dividido em duas etapas:

***Etapa de Seleção de palavras e atribuição de peso***

Nesta etapa o sistema trata os artigos do PubMed como um vetor de palavras ordenadas por peso. Estas palavras são obtidas a partir da eliminação de palavras irrelevantes para o sistema classificadas como stopwords.

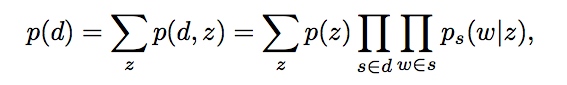
As stopwords são obtidas por duas estratégias distintas, a primeira consiste na geração do DF e TF-IDF das palavras oriundas de uma porção aleatória de artigos da base de dados do PubMed. As palavras com alto DF (número de documentos onde a palavra aparece) ou com baixo TF-IDF (métrica que informa quanto a palavra é relevante para o documento) são consideradas stopwords. A segunda estratégia consiste em considerar como stopwords as palavras que respeitam as seguintes regras: 1) palavras com menos de 3 letras; 2) palavras sem caracteres alfabéticos; 3) palavras que aparecem no **Journal of Business Research** de Jan. 2005 a 2006.

Após a eliminação das stopwords é dado um peso para cada palavra restante do documento. Este peso é obtido pela verificação da distribuição da palavra pelo documento (TF).

***Etapa de Geração do modelo probabilístico***

As palavras selecionadas na etapa anterior são usadas para treinar um classificador probabilístico que será usado para gerar uma métrica de recomendação para os novos artigos. A função utilizada para computar o grau de recomendação de um artigo é dada pela fórmula abaixo:

Sendo d um artigo, z a variável correspondente ao cluster, s um campo existente na estrutura do artigo exe., título e w uma palavra do artigo temos:



Função de Recomendação

Em seguida os autores treinam os parâmetros de probabilidade p(z) e ps(w|z) a partir dos artigos preferidos do usuário utilizando o algoritmo de Maximização de Expectativa (EM).

**Recuperação diária de novos artigos da base PubMed.**

O sistema PURE diariamente executa uma operação de recuperação dos novos artigos publicados na base de dados do PubMed. Os novos artigos são armazenados na base de dados do sistema para serem classificados de acordo com as preferências de cada usuário.

**Recomendação dos artigos**

Para cada artigo recuperado, são extraídas palavras que são usadas como base para geração do critério de ordenação que é dado de acordo com a função de recomendação definida anteriormente. Como forma de ajustar possíveis desvios do algoritmo de recomendação, os autores apresentam um score adicional para cada artigo. O Z-score é obtido através do agrupamento dos artigos em conjuntos de artigos com o mesmo número de palavras. Deste modo é separado o mínimo e o máximo graus de recomendação do grupo onde o Z-score do artigo é dado pela formula:



Os artigos com o maior Z-score são então recomendados para o usuário.

1. Estado da Arte

**Sistemas de recomendação**

Sistemas de recomendação são softwares especializados em apresentar “opções” para serem usadas por seus usuários [10,11]. Esses sistemas auxiliam pessoas que não possuem muita experiência ou competência para pesquisar dados acerca de um determinado assunto.

Em linhas gerais, os sistemas de recomendação procuram oferecer as melhores opções de resposta para as necessidades dos usuários num processo de tomada de decisão.

Os sistemas de recomendação podem oferecer sugestões em diferentes domínios como por exemplo: que item comprar em uma loja virtual, que artigo ler em um site de notícias ou mesmo que restaurante visitar em uma cidade.

No sistema de recomendação do **IMDB** (**Internet Movie Database**), quando o usuário seleciona um filme do catálogo para ler a respeito, o sistema apresenta também uma lista de sugestões de filmes relacionados para o usuário.

Assim como o **IMDB**, o site de compras **Amazon** apresenta para cada produto selecionado pelo usuário, uma lista de recomendações de outros produtos que possam servir para o usuário dentro de uma mesma compra.

Sistemas de recomendação são usualmente personalizados com base em características individuais ou coletivas, todavia, os sistemas de recomendação não personalizados também tem o seu espaço. Sistemas de recomendação não personalizados são mais simples de serem implementados e geralmente são usados para a recomendações mais gerais como por exemplo, a lista dos “10 mais” de um determinado assunto ou tema.

Sistemas de recomendação personalizados tentam antecipar as necessidades do usuário analisando informações do seu perfil e levando em consideração as restrições de domínio para então recomendar sugestões. Ainda no site de compras da **Amazon**, uma vez realizada uma compra, as informações são armazenadas no perfil do usuário de modo que para as compras futuras o usuário receba como recomendação, não apenas itens relacionados pela categoria em que se encaixam, mas também por que estão relacionados com a última compra feita por ele.

**Modelos de sistemas de recomendação**

Sistemas de recomendação variam de acordo com a técnica utilizada para descobrir as preferencias do usuário. Entre os sistemas de recomendação mais comuns temos:

*Baseados em conteúdo*

Sistemas de recomendação baseados em conteúdo tentam recomendar opções que são similares a algum item que o usuário já selecionou no passado. Por exemplo, se um usuário leu um artigo sobre política em um portal de notícias, este portal poderia sugerir, em consultas futuras, artigos relacionados a política.

*Baseado em filtragem colaborativa*

Sistemas de recomendação baseados em filtragem colaborativa, tentam recomendar ao usuário opções que foram utilizadas por usuários com os mesmos interesses. Por exemplo, se um grupo de usuários selecionam o produto X e em seguida selecionam o produto Y, o sistema de recomendação entende que usuários que acessam o produto X também acessam o produto Y e passa a recomendar o produto Y toda vez que algum novo usuário se interessar pelo produto X.

*Baseado em nichos demográficos*

Este tipo de sistema de recomendação baseia-se no perfil demográfico do usuário para recomendar opções. Usuários de um determinado estado recebem recomendações diferentes de usuários de outros estados. Os nichos demográficos podem ser: idade, sexo, língua etc.

*Baseado em conhecimento*

Sistemas de recomendação baseados em conhecimento tentam recomendar ao usuário opções baseando-se no conhecimento específico do domínio do sistema. Este modelo também é reconhecido como um modelo baseado em caso onde o problema é a analise das necessidades do usuário e a solução é o conjunto de opções a serem recomendadas.

Sistemas de recomendação com esta técnica tendem a trabalhar melhor que os outros no início, porém se não são acompanhados de componente de aprendizagem se tornam ineficientes.

Neste sistema o principio de recomendação está baseado na similaridade de uma solução para um dado problema.

*Baseado na comunidade*

Este tipo de sistema de recomendação procura basear-se nas relações que o usuário possui com sua rede de amigos para realizar as recomendações. Neste modelo acredita-se que a recomendação de pessoas ligadas ao usuário tendem a ser mais efetivas. Esta abordagem tem se tornado bastante atrativa tendo em vista o grande crescimento das redes de relacionamento.

*Sistemas híbridos de recomendação*

Sistemas de recomendação híbridos procuram utilizar técnicas de outros sistemas de recomendação de forma a complementar as deficiências entre as técnicas.

1. Proposta

O objetivo da proposta é mostrar uma alternativa automática para auxiliar os jornalistas no relacionamento de matérias. O sistema de publicação de matéria já está desenvolvido e será sobre ele que o sistema especialista deverá atuar. Para entender como o sistema especialista vai atuar é importante explicar o funcionamento do sistema de publicação de matérias do techtudo.

* 1. Sistema de Publicação de Matérias

A produção de matérias do techtudo tem uma equipe com 10 jornalistas divididos em aproximadamente 7 editorias, com uma produção diária de 40 novas matérias.

O sistema foi concebido sob plataforma web de modo que pode ser acessível pela intranet coorporativa.

Os editores têm acesso a uma interface de publicação de matéria. Esta interface apresenta um formulário com os campos da estrutura de uma matéria (título, subtítulo e corpo), após editada a matéria, em um campo de seleção, o editor seleciona a editoria a qual a matéria pertence e em seguida adiciona o elemento saiba mais, responsável pela criação do elemento visual de matérias relacionadas. Este elemento permite que o usuário adicione manualmente links para matérias previamente cadastradas no sistema.

Uma vez publicada a matéria, esta passa a ser acessível pela página da categoria a qual foi associada.

Ex.ª: uma matéria da editoria jogos:

<http://www.techtudo.com.br/jogos/noticia/2011/09/vem-ai-o-rpg-online-de-power-rangers.html>

* 1. Sistema especialista

O sistema especialista deve ser capaz de sugerir, após a edição da matéria, uma lista de links para matérias relacionadas de modo que o editor não precise fazer esta operação manualmente.

O sistema deve permitir que o editor substitua os links de forma manual caso não concordem com a recomendação fornecida.

* 1. Verificação

A verificação é a etapa de avaliação dos resultados obtidos com a aplicação da solução sugerida para um determinado problema objetivando a aprovação ou rejeição do seu uso.

* + 1. Avaliação da precisão do algoritmo

Para avaliar a contribuição esperada pelo algoritmo, propomos uma comparação baseada na probabilidade de uma matéria ser recomendada uma vez que ela foi recomenda anteriormente por matérias similares.

Para montagem desta avalição separamos uma matriz de colaboração item-item (Tabela 1) onde em um eixo temos as matérias recomendadas e no outro, temos as matérias vistas em seguida pelo usuário.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Matéria A | Matéria B | Matéria C | Matéria D |
| Matéria A | 0 | 0 | 35 | 0 |
| Matéria B | 40 | 0 | 0 | 4 |
| Matéria C | 0 | 5 | 0 | 23 |
| Matéria D | 0 | 3 | 23 | 0 |

Os dados para montagem da matriz foram obtidos junto ao sistema de monitoração de navegação dos usuários, o **google** **analitycs**.

Como pode ser visto na tabela 1, após consumir a matéria A os usuários consumiram a matéria C e não passaram nas matérias B ou D.

Para o modelo comparativos vamos obter as matérias que os usuário viram antes da matéria recomendada e verificar qual a probabilidade da matéria nova pertencer a este grupo.

* 1. Informação sobre o corpus utilizado

O corpus de trabalho conta com um conjunto de matérias organizadas em categorias, conforme exposto na tabela tabela 1:

|  |  |
| --- | --- |
| Categoria | Total |
| Artigos | 1069 |
| Dicas e Tutoriais | 1616 |
| Notícias | 1286 |
| Reviews | 629 |
| Jogos | 2309 |
| Rio de Janeiro | 1268 |
| Curiosidades | 144 |
| Humor | 139 |
|  | 8460 |

Tabela 1 - Total de artigos por categoria

A matéria possui uma estrutura de dados básica, contudo, apenas os campos título, subtítulo, corpo e categoria serão utilizados neste trabalho.

Observações quanto ao corpus:

1 – Cerca de 80% do corpus possui recomendações realizadas manualmente, deste modo podemos realizar a verificação proposta.

2 – Não temos com esse projeto a expectativa de um percentual de acertos de sugestões de 100% até porque, a classificação humana também tem seu percentual de acertos e erros se forem observadas avaliações entre pessoas diferentes. Portanto, a avaliação da viabilidade do algoritmo será dada em função da aprovação da equipe de editores.

1. Esboço do documento final

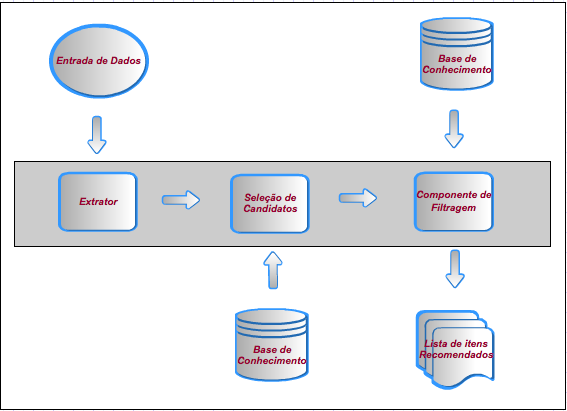
Esta seção apenas descreve a proposta de organização da dissertação final.

1. Introdução
2. Trabalhos relacionados
3. Estado da arte
4. Resultados dos experimentos
5. Conclusão
6. Referências

1. Plano de ação (esboço da solução)

Já foram feitos alguns experimentos com o corpus que será utilizado. A implementação inicial tem o seguinte fluxo:

**Arquitetura do sistema**



1. **Entrada de dados**

A etapa de entrada de dados é o ponto de partida para o sistema de recomendação proposto, nesta fase, é apresentado ao editor uma interface web no formato de um formulário que solicita ao editor um conjunto de informações: Estas informações podem ser divididas em obrigatórias e não obrigatórias segundo a tabela 1.

As informações obrigatórias são necessárias para constituir uma notícia que pode ser publicada.

|  |  |  |
| --- | --- | --- |
| Título | O título da notícia é um dado textual limitado em 255 caracteres alfa numéricos. | Obrigatória |
| Subtítulo | O subtítulo da notícia é um dado textual limitado em 1000 caracteres alfa numéricos. | Obrigatória |
| Corpo | O corpo da notícia é um dado textual com um limite de caracteres alfa numéricos grande. O Editor pode fazer uso de markup html na construção do conteúdo da notícia | Obrigatória |
| Editoria Principal | Dado que agrupa o conjunto de notícias. Esta informação é selecionada na interface através de um caixa de seleção. Este agrupamento é criado pelos editores e geralmente refletem seções do site na internet. Exemplo de Editoria para o techtudo: Jogos, Dicas e Tutoriais, Reviews | Obrigatória |
| Editorias Secundárias | Agrupamentos secundários para organização das notícias | Não Obrigatória |
| Entidades Associadas | Informações extras que associam as notícias | Não Obrigatória |

1. **Extrator**

Nesta etapa o objetivo é construir um entendimento acerca das informações recebidas do editor na etapa de entrada de informações. Para isso, é necessário um tratamento das informações com técnicas de mineração de textos para extrair o máximo de informações relevantes.

* Fase de limpeza:

O corpo da matéria é construído com texto livre o que permite aos editores a inclusão de mark-ups html para marcação do texto, com isso é comum termos palavras em negrito, itálico ou até mesmo elementos mais elaborados com tabelas e div’s que agrupam dados textuais. O extrator elimina os caracteres que representam marcadores html e realiza a conversão de normas de acentuação típicas da notação html para a tabela ascii. Assim temos:

Ex. <strong> OL&Aacute; MUNDO! </Strong>

É convertido para OLÁ MUNDO!

* Fase de geração do Bag of Words:

Com o texto livre de caracteres de marcação, o extrator realiza a eliminação de todos os caracteres de pontuação e realiza a quebra do texto em palavras utilizando como critério a presença de espaço em branco entre elas.

* Fase de geração do TF:

As palavras são então contadas e ordenadas em função de sua presença no texto, as palavras consideradas como stopwords e as palavras classificadas gramaticalmente como verbo são eliminadas. Assim podemos descobrir quais palavras são realmente relevantes para o texto.

Para remover as palavras chamadas stopwords, o extrator faz uso de um dicionário de palavras marcadas como stopwords. Este dicionário foi gerado com a recuperação das pala lavras com o menor TF-IDF que significa o grau de relevância da palavra para todo o corpus.

Para remover todas as palavras que gramaticalmente são classificadas como verbo, o extrator faz uso de um tagger gerado com base em uma acervo de artigos da folha de SP.

Após o tratamento do texto são também separados as entidades fornecidas na criação da matéria. Elas são armazenadas em uma array de entidades. Este mesmo procedimento é realizado para a informação de editoria.

1. **Seleção de Candidatos**

Nesta etapa é avaliado pesos e critérios para realização de consultas no corpus. Uma vez analisada a notícia e extraído os dados necessários poderemos estabelecer caminhos de procura de informações.

O corpus está armazenado em um servidor de busca por texto baseado em **lucene**, de tal forma que as consultas são realizadas por palavras chave e levam em consideração algumas regras de ordenação:

* As matérias são ordenadas por data de publicação das mais recentes para as mais antigas.
* Palavras chaves encontradas no campo “título” tem o dobro da relevância quando encontrada no corpo do texto.
* São observadas matérias que foram publicadas dentro de no máximo, 60 dias anteriores a data atual
* São selecionados um número fixo de matérias por consulta para permitir que as consultas sejam rápidas e padronizadas.

Embora tenhamos uma infinidade de heurísticas possíveis para seleção dos candidatos, optamos pela elaboração de 3 heurísticas iniciais com os seguintes pesos:

1. Mais da mesma entidade (peso 5)

Esta regra determina que o sistema selecione candidatos que falam da mesma entidade fornecida pelo usuário.

1. Mais da mesma editoria (peso 2)

Esta regra determina que o sistema selecione candidatos que participam da mesma editoria fornecida pelo usuário.

1. Mais das palavras mais relevantes (peso 1)

Esta regra determina que o sistema utilize a distribuição de frequência de termos e realize consultas por matérias que se relacionem com o termo mais relevante para o texto.

1. **Componente de Filtragem**

Nesta etapa é observada a probabilidade de um item sugerido ser uma boa sugestão para uma determinada matéria.

Para realizar essa atividade o componente de filtragem faz uso de um classificador que avalia quão similar é a matéria ao conjunto de matérias que foram vistas no momento anterior a matéria sugerida. Para tanto, construímos a matriz item-item que representa um fluxo de origem para destino.

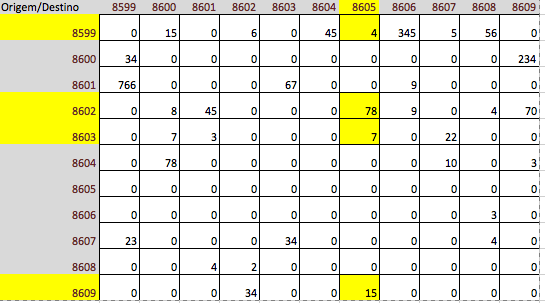


Tabela 3

Segundo a tabela 3 quem viu a matéria 8605, viu antes o conjunto de matérias 8599, 8602, 8603, 8609.

Desta forma poderíamos dizer que, se uma nova matéria fizesse parte do conjunto de matérias visitadas anteriormente a matéria 8599, poderíamos sugerir a matéria 8599 como matéria recomendada a essa nova matéria.

Para avaliar se a nova matéria faz parte do conjunto de matérias visitadas anteriormente, o sistema faz uso de um classificador binário que aponta a probabilidade da matéria pertencer ao grupo.

Após a avaliação da probabilidade o sistema seleciona entre as matérias candidatas, aquelas que apresentaram a maior probabilidade de serem recomendadas.

1. Cronograma

Nesta seção é apresentado apenas um cronograma especificando as tarefas que deverão ser feitas para a conclusão do trabalho de dissertação.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Jul | Ago | Set | Out | Nov | Dez |
| Proposta | X | X |  |  |  |  |
| Aperfeiçoamento do pré-processamento com uso de lematização |  | X |  |  |  |  |
| Estratégia de seleções de termos para consulta |  | X | X |  |  |  |
| Aperfeiçoamento da etapa de recuperação dos resultados |  |  | X |  |  |  |
| Execução e Testes do Algoritmo |  |  | X |  |  |  |
| Escrita do capítulo sobre Resultados |  |  |  | X |  |  |
| Finalização da dissertação |  |  |  |  | X | X |

Referências Bibliográficas

1 - [BOFENG ZHANG, XIN XU, JINSHU SU - 2007]. **An Ensemble Method for Multi-class and Multi-label Text Categorization:** National University of Defense Technology.

2 - [Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze - 2008] **Introduction to Information Retrieval**, Cambridge University Press.

3 - [Min-Ling Zhang, Jose M. Penã and Victor Robles – 2009]. **Feature Selection for Multi-Label Naive Bayes Classification**: College of Computer and Infor- mation Engineering, Hohai University, Nanjing 210098, China;

4 - [Arni Darliani Asy’arie, Adi Wahyu Pribadi – 2009]. **Automatic News Articles Classification in Indonesian Language by Using Naive Bayes Classifier Method**: Universitas Islam Negeri Syarif Hidayatullah Jakarta, Universitas Pancasila Jl. Srengseng Sawah, Jagakarsa Jakarta.

5 - [Chin-Yew Lin, Eduard Hovy – 1997]. **Identifying Topics by Position,** Information Sciences Institute of the University of Southern California

6 - [Dipanjan Das, Andre F.T. Martins - 2007] **A Survey on Automatic Text Summarization,** Language Technologies Institute Carnegie Mellon University.

7 - [Jacob Perkins - 2010]. **Python Text Processing with NLTK 2.0 Cookbook**.

8 -[LUIZ CLÁUDIO GOMES MAIA – 2008] **Uso de Sintagmas Nominais na Classificação Automática de Documentos Eletrônicos**. UFMG.

9 -[Bruno Magalhães Nogueira - 2009]. **Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos**. ICMC-USP.

10 -[Francesco Ricci, Lior Rokach, Bracha Shapira - 2010]. **Introduction to Recommender Systems Handbook**.

11 -[Bracha Shapira, Lior Rokach - 2010]. **Recommender Systems and Search Engines – Two sides of the same Coin!?**. Department of Information Systems Engineering, Ben-Gurion University.

12 -[TAKASHI YONEYA , HIROSHI MAMITSUKA1 - 2007] **PURE: A Pubmed Article Rcommendation System Based on Content-Based Filtering.** Bioinformatics Center, Kyoto University, Gokasho Uji, Japan and Discovery Research Laboratories, Kirin Pharma, Miyahara, Takasaki, Japan

13 – [WORLD WIDE WEB SIZE] **Daily Estimated Size**, 2011 Disponivel em: <http://www.worldwidewebsize.com/>. Acesso em: 08 Jul 2011.

14 – [MINIWATTS MARKETING GROUP] **Internet World Stats**, 2011. Disponivel em: <http://www.internetworldstats.com/stats.htm>. Acesso em: 19 Jan 2011.

15 – [Greg Linden, Brent Smith, Jeremy York] **Amazon.com Recommendations Item-to-Item Collaborative Filtering**, 2003 - IEEE INTERNET COMPUTING.

16 - [James Davidson, Benjamin Liebald, Junning Liu] - **The YouTube Video Recommendation System**, 2010, Google Inc

17 - [Ying Huang] - **An Intelligent adaptative news filtering system**, 2001, University of Science & Technology of China.

18 - [Abhinandan Das, Mayur Datar Google,] **Google News Personalization: Scalable Online Collaborative Filtering**, 2007, Google Inc