INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
UNIVERSIDADE DE SÃO PAULO

Proposta para monografia

MAC 499 - Trabalho de formatura supervisionado

Marcela Ortega Garcia - n° USP 5638866

1 Recuperação de Informações em Banco de Dados Textuais

Aluno: Marcela Ortega Garcia
Supervisor: Prof. Dr. João Eduardo Ferreira

2 Resumo

Um dos principais objetivos da área de banco de dados é armazenar e recuperar dados de maneira eficiente. Para possibilitar acesso rápido ao conteúdo armazenado, sistemas de banco de dados utilizam estruturas conhecidas como "índices". O crescimento do uso de bancos para dados sem estrutura definida desencadeou a necessidade de técnicas de indexação diferenciadas. Dentre esses tipos de dados, encontram-se os textos e um exemplo que ilustra esse cenário é a pesquisa em páginas *Web*, um conjunto volumoso corriqueiramente consultado.

Uma importante diferença entre banco de dados de textos e os convencionais é a maneira como são acessados. Enquanto consultas em banco de dados comuns são expressões lógicas exatas como "Quais estudantes do IME estão matriculados em MAC499?", pesquisas em banco de dados textuais são inexatas como "Quais assuntos os alunos matriculados em MAC499 no ano de 2009 escolheram abordar no trabalho de formatura?". Usualmente não é possível traduzir uma pergunta em uma expressão lógica.

Como não há um mecanismo exato de determinar se um documento é uma resposta para a pergunta, consultas a banco de dados textuais devem ser capazes de identificar resultados relevantes, pertinentes à consulta. Para isso, existem diversas técnicas de indexação como "Arquivos invertidos", "Assinatura de arquivos" e "Vetores de sufixos" e a eficiência de cada uma é medida pela proporção de documentos relevantes que são encontrados.

Com o objetivo de aumentar a qualidade das respostas, outra maneira de pesquisa, conhecida como "Busca Semântica", começou a ser desenvolvida. A idéia principal é utilizar a semântica, ou seja, o significado das palavras, para encontrar respostas relevantes. Enquanto pesquisas convencionais procuram por palavras chaves que se igualam às palavras da consulta, a busca semântica é um conjunto de técnicas para obter conhecimento de um volume rico de dados.

3 Objetivos

O objetivo deste trabalho de formatura é fazer um estudo aprofundado de técnicas de indexação em bancos de dados textuais. Além de estudar técnicas e algoritmos disponíveis em conceituados livros da área, pretendemos analisar a tecnologia de busca utilizada pelo Google e também as técnicas utilizadas pelo Lucene, uma biblioteca de recuperação de informação.

Após adquirirmos um conhecimento teórico, pretendemos utilizar o Lucene para implementar um módulo no sistema do Centro de Estudos do Genoma Humano (CEGH) da USP. Esse sistema armazena anotações de médicos sobre pacientes e, inicialmente, a idéia é realizar as consultas sobre esses dados.

Pretendemos também iniciar um estudo sobre "Busca Semântica", uma área de desenvolvimento recente que achamos bem interessante e promissora. Queremos direcionar esse estudo no entendimento de como funcionam os recém lançados "WolframAlpha" e "Google Squared".

4 Atividades já realizadas

As atividades realizadas até o momento foram:
- Levantamento bibliográfico de referências sobre recuperação de informações;
- Início do estudo de técnicas de indexação;
- Análise dos dados disponíveis no banco do Centro de Estudos do Genoma Humano da USP;
- Pesquisas sobre busca semântica.

5 Cronograma de atividades

Atividade/Mês	Jun	Jul	Ago	Set	Out	Nov
Estudo das técnicas de indexação	X	X	X			
Análise das técnicas utilizadas pelo Google e pelo Lucene			X			
Implementação do módulo para o sistema do CEGH				X	X	
Estudo sobre busca semântica					X	X
Elaboração da monografia		X	X	X	X	X
Elaboração do pôster e apresentação					X	X

6 Estrutura esperada da monografia

A monografia terá uma parte objetiva e uma parte subjetiva, seguindo a estrutura proposta na página da disciplina:

Parte objetiva

1. Introdução: importância da recuperação de informações e os objetivos do trabalho.
2. Conceitos e tecnologias estudadas: conceitos de recuperação de informações, técnicas de indexação e tecnologias estudadas.
3. Atividades realizadas e resultados obtidos: análise dos resultados obtidos, relacionando a teoria com o sistema do CEGH e detalhes da implementação do módulo.
4. Conclusões
5. Bibliografia

Parte subjetiva

1. Experiência e aprendizando elaborando um TCC.
2. Desafios e frustrações encontrados.
3. Disciplinas cursadas no BCC mais relevantes para o trabalho.
4. Trabalhos futuros.

7 Referências

1. Baeza-Yates R.; Ribeiro-Neto B. *Modern Information Retrieval*. Addison Wesley, 1999.
2. Bertino E.; OOI B.; Sacks-Davis R.; Tan K.; Zobel J.; Shidlovsky B.; Catania B. *Indexing Techniques for Advanced Database Systems*. KAP, 1997
3. Lucene - http://lucene.apache.org
4. WolframAlpha - http://www.wolframalpha.com
5. Google Squared - http://www.google.com/squared