Integrando recuperação de informação em banco de dados com Hibernate Search

IME - Instituto de Matemática e Estatística

Alunos: Gustavo Kendi Tsuji e Leonardo Tadashi Kamaura
Orientador: João Eduardo Ferreira

Universidade de São Paulo USP

Introdução

Sistemas com grande volume de dados apresentam um grande revés: a consulta no banco de dados são altamente custosas e lentas, o que muitas vezes reflete no funcionamento do restante do sistema.

Um bom exemplo é o caso de sistemas de buscas na internet. Segundo dados, a partir de 2000, estima-se que a internet tenha um crescimento de 65% ao ano [1]. Além de lidar com um enorme banco de dados, é preciso encontrar soluções que viabilizem recuperar informações relevantes de forma rápida e eficiente.

Busca indexada

A busca indexada funciona como uma pesquisa de um termo num grande índice remissivo. Basta consultar os índices para saber a localização dos registros no banco de dados. Devido a sua estrutura, buscas em índices são mais eficientes do que consultas diretas no banco de dados.

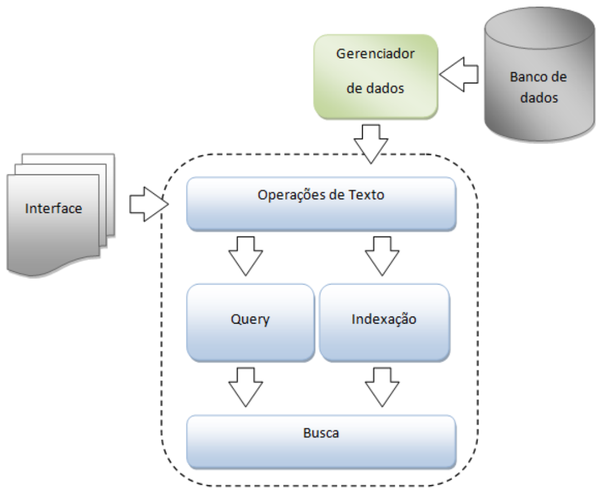

Fig. 1 - Esquema geral de um sistema de RI

Modelo Vetorial

O modelo vetorial é um modelo algébrico para representar documentos como vetores. Para cada termo que ocorre num documento, atribui-se um peso. A fórmula mais comum baseia-se na freqüência dos termos no banco de dados (tf-idf), que serve para estimar o grau de similaridade dos registros. A fórmula (fig.2) corresponde a distância do coseno entre a query q e o documento dj.

$$sim(d_j \cdot q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|},$$

Fig. 2 - Cálculo do grau de similadridade

Projeto Colméia

O projeto colméia [3] é um sistema de gerenciamento de bibliotecas e possui um banco de dados estruturalmente complexo. Seu acervo contempla aproximadamente 40000 livros. Utilizamos este banco como estudo de caso.

Ferramentas

O Lucene[4] disponibiliza uma interface que possibilita a indexação de dados, além de fornecer recursos para executar uma busca indexada. Já o Hibernate Search[5] é um módulo do arcabouço Hibernate que integra o Lucene com o banco de dados. É adequado principalmente em projetos orientados a objetos, encapsulando entidades e relaciomentos em classes. O objetivo é simplificar e deixar transparente a maneira como a aplicação lida com os dados do banco.

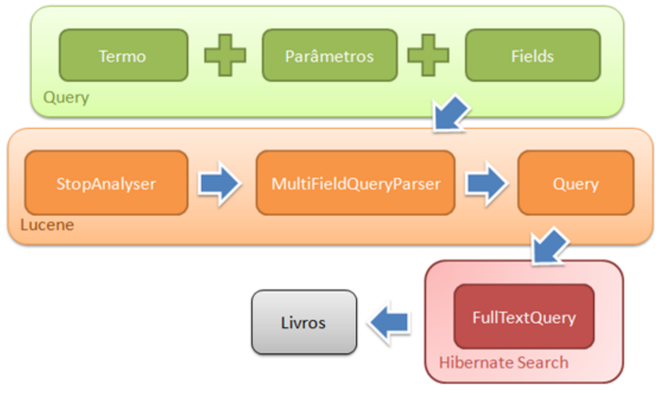

Fig. 3 - Diagrama de integração do Hibernate e do Lucene

Resultados

Com a aplicação dos estudos sobre os arcabouços, foi possível implementar a busca indexada sobre as obras do acervo da biblioteca do IME. Um exemplo de uma possível consulta: localizar todos os livros cujo assunto seja "banco de dados" ou que contenham o termo "database" no título da obra, mas que foram publicadas entre os anos de 2002 e 2008.

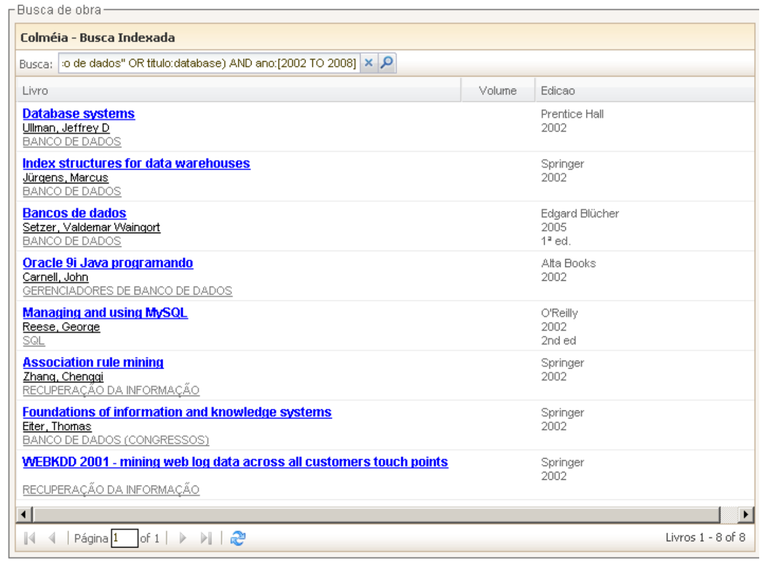

Fig. 4 - Resultados da consulta no Colméia

A base de cálculo para o ranqueamento segue na tabela abaixo:

Query: (assunto:"banco de dados" OR titulo:database) AND ano:[2003 TO 2008]
R1: Database systems
R2: Index structures for data warehouses
R8: WEBKDD 2001 - mining web log data across all customers touch points

			Banco de dados			Database			2003 - 2008
	Score	Coord	Weight	Query Weight	Field Weight	Weight	Query Weight	Field Weight	Constant Score Query
R1	7.505524	1	5.2418785	0.8634171	6.071085	2.1925368	0.4994541	4.3898664	**0.07110896**
R2	2.6920483	0,5	5.2418785	0.8634171	6.071085	-	-	-	**0.07110896**
R8	1.3815786	0,5	2.6209393	0.8634171	3.0355425	-	-	-	**0.07110896**

Fig. 5 - Cálculo detalhado dos resultados

Referências Bibliográficas

[1] WORLD Internet Usage Statistics News and World population Stats. Disponível em : <http://www.internetworldstats.com/stats.htm>. Acesso em: 6 nov. 2008
[2] BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. Modern Information Retrieval. Addison Wesley, 1999
[3] REPOSITÓRIO do Colméia. Disponível em: <http://colmeia.incubadora.fapesp.br>. Acesso em: 6 nov. 2008
[4] GOSPODNETIC, Otis; HATCHER, Erik. Lucene in Action. Manning, 2005
[5] HIBERNATE Search. Disponível em: <http://www.hibernate.org/hib_docs/search/reference/en/html/>. Acesso em: 6 nov. 2008

HIBERNATE Lucene Ext JS