Gerador Automático de Arquivos HTML de Ajuda para Cursos a Distância via Internet

Dr. Dilvan de Abreu Moreira

Paulo Sérgio Salla Sá

Instituto de Ciências Matemáticas e de Computação – ICMC - USP
Departamento de Ciência de Computação e Estatística
Av. Dr. Carlos Botelho, 1465 – Caixa Postal –668
CEP: 13560-970 – São Carlos - SP
dilvan@icmc.sc.usp.br

Keywords: Distant Education, Search Engines, Software Agents, Java.

Sub-área: Aplicações multimídia e hipermídia para Web.

O Gerador Automático de Arquivo HTML de Ajuda (GAAHA) é um programa para a geração de arquivos de indexação e ajuda em HTML baseado na análise automática de documentos. Arquivos de indexação e ajuda têm uma dinâmica muito rápida de atualizações de seu conteúdo. Este programa focaliza principalmente: a geração e manutenção de arquivos de ajuda para documentos de cursos para educação a distância (EAD) via Internet e a criação de um formato de arquivos de ajuda (Help files) eficiente, para facilitar e agilizar a pesquisa de tópicos ou palavras desses cursos. Uma técnica de seleção de palavras significantes é empregada no momento da indexação, com o objetivo de armazenar apenas palavras que realmente tenham um valor significativo.

O aluno, que participa de um treinamento ou curso de educação a distância, poderá aproveitar o sistema de ajuda para pesquisa sobre tópicos relacionados ao curso ou treinamento em questão.

O GAAHA tem um robô *parser*, responsável pela geração de uma base de dados indexada, que é utilizada para pesquisas de palavras-chave entradas pelo usuário e também para geração de uma vista hierárquica de todos os *links* de um curso, aqui chamado de mapa do curso. Este sistema é composto por três subsistemas: um agente *crawler* para a indexação das informações; um agente *mentor* que é uma rotina de pesquisa executada no servidor, responsável pela pesquisa no banco de dados; e um cliente WWW para a entrada dos dados de pesquisa.

O agente *crawler* e o agente *mentor* são os dois módulos principais desenvolvidos neste projeto. O cliente consiste de um *Web Browser* (tal como o Netscape ou Explorer) e um documento HTML criado dinamicamente pelo agente *mentor*. O agente *crawler* indexa cursos de EAD, todos os documentos que compõem este curso são indexados sem qualquer interação externa do administrador. Basicamente, o programa gera tabelas em um banco de dados relacional com as palavras mais significativas, os *links* de referências, os *headers* e os títulos das páginas HTML. Cada informação útil gerada é relacionada com os locais onde elas aparecem.

O tipo de indexação aplicada neste projeto para a análise da estrutura de um texto é a indexação automática conhecida como *full-text* (Yates, 1996). Ela indexa

automaticamente todas as palavras do documento. Para diminuir o problema do tamanho dos índices e também para armazenar apenas palavras significativas, um filtro foi implementado. Para a implementação deste filtro, foi utilizada uma técnica estatística de indexação automática, baseando-se na freqüência de ocorrência de palavras nos documentos (Chen et al., 1996).

O resultado da indexação é a construção dinâmica, pelo agente *mentor*, de uma página HTML de ajuda, correspondente ao curso em questão, como mostrado na figura 1. Esta página contém basicamente um botão que possibilita a construção dinâmica do mapa do curso, um campo para digitação das palavras-chave de pesquisa ao banco de dados e um *link* para uma página informativa contendo informações de utilização da máquina de pesquisa.

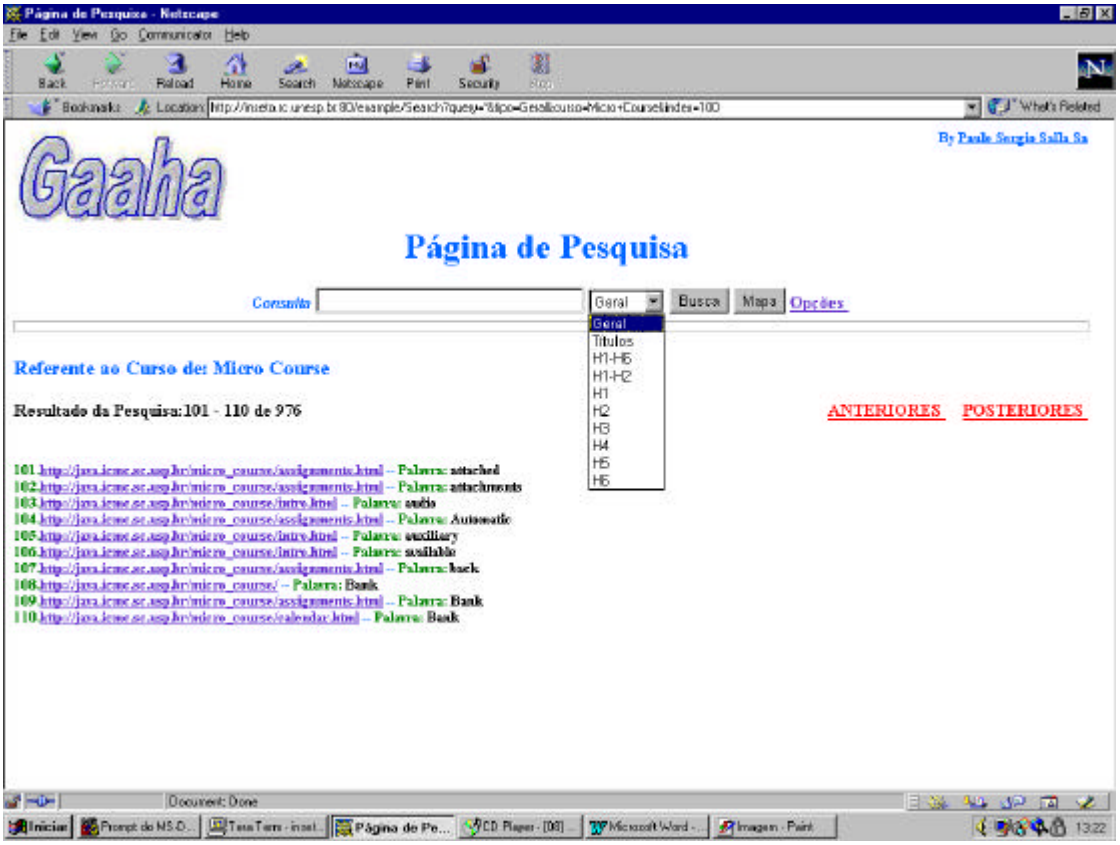

Figura 1 – Execução do agente mentor.

Referências

R. B. Yates, "An extended model for full text databases", *Journal of the Brazilian Computer Society*, v.2, n.3, April 1996.

H. Chen et all, "A concept space approach to addressing the vocabulary problem in scientific information retrieval", *Experiment on the Worm Community System*, MIS Department, University of Arizona, July 1996. (http://ai.bpa.arizona.edu/ papers, 1998).