Arquitetura de um Mecanismo para Busca Especializada na WEB

FERNANDO RENIER GIBOTTI[1]
GILBERTO CÂMARA[2]
RENATO ALMEIDA NOGUEIRA[3]
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

Palavras-chave
mecanismo de busca - dados geográficos - web.

Resumo

O desenvolvimento acelerado da Internet e o aumento de conteúdos digitais disponíveis conduziram o desenvolvimento de mecanismos de busca que facilitassem a recuperação de informações na Web. Entretanto, estes mecanismos apresentam limitações principalmente quando se trata da recuperação de conteúdos especializados. Dados geográficos não são encontrados na Web por mecanismos convencionais, pois tais mecanismos não estão preparados para evidenciar este conteúdo. Nesse contexto, este trabalho apresenta uma arquitetura para acesso e recuperação de dados geográficos na Web, discorrendo sobre suas principais características, arquitetura, tecnologias envolvidas e performance.

1. Introdução

O mundo é um lugar imperfeito e imprevisível. Desde sua origem no final da década de 60, a Internet cresceu rapidamente e passou de um projeto de pesquisa a uma vasta coleção de documentos heterogêneos. A Web é uma rede composta por bilhões de páginas multimídia interligadas (imagens, sons, textos, animações, arquivos etc) desenvolvidas de forma descoordenada por milhões de pessoas. O aumento expressivo na quantidade de informações publicadas e a falta de estruturação do conteúdo de documentos disponíveis constituem um desafio para a recuperação de dados.

Mecanismos de busca como Google (www.google.com), Altavista (www.altavista.com), Yahoo (www.yahoo.com) surgiram com o objetivo de melhorar a procura e análise de informações na Web [Glover 2002]. Entretanto, estes mecanismos possuem algumas limitações, que variam da cobertura dos documentos indexáveis na Web [Lawrence and Giles 1998] à quantidade de respostas indesejadas retornadas por uma pesquisa, dificultando o acesso à informação. Para superar essas limitações pesquisadores estão empreendendo grandes esforços para melhorar a qualidade de informações semânticas nas páginas Web.

Seguindo Berners-Lee em seu paper "The Semantic Web" [Berners-Lee, Hendler and Lassila 2001] e o interesse em ontologias [Gruber 1995] [Guarino and Giaretta 1995] [Wiederhold 1994], a comunidade de TI está trabalhando em propostas para organizar a informação visando a interoperabilidade e a fácil recuperação. Nestas propostas estão incluídas linguagens tais como OWL [Masolo, Gangemi, Guarino et al. 2002]. A comunidade geoespacial também participa desses esforços [Egenhofer 2002] e propostas como a linguagem GML auxilia fornecendo padrões para a interoperabilidade de dados espaciais.

Entretanto, a utilização de linguagens tais como OWL ou GML requer maiores esforços dos produtores de dados. Produzir e organizar ontologias são uma tarefa emergente, que pode estar além das habilidades e possibilidades de muitas organizações. Uma alternativa para padronizar a distribuição de dados é fornecer

[1] gibotti@dpi.inpe.br
[2] gilberto@dpi.inpe.br
[3] renato@faimi.edu.br

mecanismos de busca especializados. Estes mecanismos sabem que algumas comunidades produzem documentos estruturados e que a partir de hipóteses muitas informações semânticas podem ser diretamente recuperadas. Os exemplos mais recentes são mecanismos para recuperar e indexar artigos científicos, tais como Google Scholar e Citesser [Giles, Bollacker and Lawrence 1998]. Estes mecanismos sabem que um artigo científico possui uma estrutura bem definida: título, lista de autores, resumo, texto e referências. Estes mecanismos fornecem informação útil e o esforço exigido dos autores é a postagem dos artigos originais em uma página Web.

E sobre os dados geográficos? A Web possui uma grande quantidade de dados geoespaciais, mas os mecanismos de busca tradicionais não são especializados para reconhecê-los, gerando um distanciamento entre estes dados e os usuários. Entretanto, dados geográficos são semi-estruturados. Muitos arquivos compartilham características comuns: fornecem dados do local (no formato vetorial ou matricial) e combinam atributos. Além disso, o número de formatos para a distribuição dos dados geoespaciais é limitado. Similarmente como acontece em artigos científicos (onde o formato PDF é o predominante), produtores de dados geoespaciais usualmente distribuem seus dados em formatos que possibilitem sua leitura pelos usuários. Estes formatos incluem shapefiles, GML e dados matriciais em GeoTIFF.

Uma alternativa para as propostas de padronização semântica, tais como GML, é a utilização de um mecanismo de busca especializado em dados geoespaciais. Estes mecanismos especializados poderiam considerar a natureza semi-estruturada dos dados geográficos para executar a busca. A idéia é permitir o compartilhamento dos dados geográficos sem a necessidade de trabalho adicional com anotações semânticas. Alguns dos princípios para esses tipos de compartilhamento de dados estão descritos em Onsrud et al. [2004]. Com base nesta motivação, este artigo apresenta um mecanismo de busca especializado para acesso e recuperação de dados geográficos na web. Nossa proposta inclui uma tecnologia rápida e distribuída de rastreamento de dados geográficos.

Neste artigo, nós discutimos os desafios conceituais envolvidos no projeto de um mecanismo de busca para dados geográficos. Os principais desafios são três: (a) desenvolver algoritmos para a recuperação de informação a partir de dados geográficos; (b) projetar uma arquitetura robusta para manipular grandes volumes de dados; (c) incluir técnicas para a preservação da privacidade e restrições de direitos de cópia. O desenvolvimento de algoritmos para a recuperação da informação geográfica é apresentado na Seção 2. Na seção 3, discutimos as características do mecanismo. Na Seção 4, descrevemos a anatomia do mecanismo e apresentamos os resultados iniciais. Os aspectos de privacidade e direito de cópia discutidos em Onsrud et al. [2004] não foram tratados neste artigo.

2. Recuperação de Dados Geográficos

Mecanismos de busca tradicionais não possuem rastejadores e analisadores especializados em dados geográficos. Estes são preparados para buscar documentos hipertexto e hyperlinks. Esta seção discorrerá sobre os principais aspectos observados para a implementação de nosso mecanismo.

2.1. Encontrando dados geográficos

A primeira questão é como encontrar dados geográficos. Nossa escolha foi considerar que dados geográficos estão normalmente distribuídos em um conjunto de formatos predefinidos associados a sistemas GIS. Podemos citar como exemplo o formato shapefile criado pelo software ArcView que é uma forma de troca de dados geográficos bem aceita pela comunidade GIS. Desta forma, nosso mecanismo tenta recuperar arquivos GIS mais comumente utilizados. Em nosso protótipo atual, são recuperados arquivos shapefile, mas o mecanismo pode ser facilmente estendido para recuperar outros formatos GIS.

2.2. Obtendo significado de dados semi-estruturados

O que há de especial em dados espaciais? Nós consideramos que grande parte dos dados geográficos contém informações que permitem suposições sobre seu conteúdo semântico. O mais óbvio são as coordenadas geográficas. Por exemplo, shapefiles contêm informação sobre os limites do dado, mas não dados específicos de projeção e datum. Usualmente, nós podemos inferir tais informações. Coordenadas na projeção lat/long possuem uma gama diferente das coordenadas UTM.

O segundo tipo de conteúdo que permite suposições são os nomes de lugares. Muitos conjuntos de dados geográficos incluem nomes de lugares, ou diretamente ou por meio de indexação (tais como FIPSNO nos EUA ou IBGE no Brasil). As colunas para nomes de lugares são usualmente associadas com rótulos tais como NAME, NOME ou NAMEN. Uma vez que uma coluna com nomes de lugares é encontrada, podemos comparar seu conteúdo com gazetteers tais como USGS´s Geographic Names Information System (GNIS) (http://geonames.usgs.gov) ou a NGA's GEOnet Names Server (http://earth-info.nga.mil/gns/html/). Gazetteers são utilizados efetivamente na Alexandria Digital Library [Frew 1998].

Os dois tipos de informação acima (projeção e lugar) são suficientes para uma indexação geoespacial. Para mais detalhes, um segundo tipo de suposições é necessário. O rastreador precisa explorar os rótulos das colunas de dados e deduzir seu conteúdo utilizando informações tais como variáveis e seus valores. Por exemplo, inteiros positivos em dados geoespaciais provavelmente sejam dados contáveis. Valores reais entre -1 e +1 possivelmente são índices. Nós chamamos este processo de descoberta como data wayfinding. Na versão atual, o GeoDiscover não privilegia este processo, mas iremos inclui-lo futuramente.

2.3. Dados sobrepostos

A existência de cópias de dados geográficos em diferentes sítios da Web gera um problema para a indexação dos arquivos, pois pode introduzir duplicações na base de dados. A detecção de dados similares é possível pela análise de sua estrutura e de alguns atributos intrínsecos dos arquivos, tais como nome, data da criação, tamanha e tipo. Através destes atributos intrínsecos, nosso mecanismo identifica arquivos similares e evita o armazenamento redundante.

2.4. Classificação dos produtores de dados

Mecanismos de busca tradicionais utilizam diferentes formas para classificar os principais sítios da Web. O mecanismo Google, utiliza o método de PageRank para priorizar os resultados de busca por palavras-chave [Page, Brin, Motwani et al. 1999] [Gerhart 2002]. Similarmente ao método para classificar páginas utilizando a informação dos links, o mecanismo Citeseer [Giles, Bollacker and Lawrence 1998] classifica artigos científicos como hubs e autoridades baseado no gráfico de citações.

O objetivo de um mecanismo de busca é possibilitar às pessoas recuperar dados geográficos de forma segura quanto à qualidade e origem dos dados. O mecanismo desenvolvido identifica produtores de dados baseado em três aspectos: quantidade de dados geográficos disponível no sítio da Web; quantidade de arquivos capturados pelos usuários a partir dos arquivos disponíveis no sítio; e pela indicação se o sítio é um hub ou uma autoridade. Nesta abordagem, hubs são sítios que recomendam outros sítios que contenham dados geográficos e autoridades são sítios que são recomendados por muitos hubs.

Ao retornar uma lista de dados geográficos para o usuário, o mecanismo desenvolvido apresenta, para cada dado, seu produtor (por exemplo, IBGE), a quantidade de dados geográficos produzidos por este produtor (por exemplo, 1080 shapefiles) e a quantidade de downloads executados a partir do sítio deste produtor (por exemplo, 10.112 shapefiles requisitados pelos usuários).

3. Arquitetura do Sistema

O mecanismo de busca desenvolvido tem algumas características que auxiliam no processo de busca de dados geográficos e que melhoram significativamente os resultados retornados. Primeiro ele classifica os produtores de dados geográficos. Segundo, ele utiliza informações adicionais presentes nos hyperlinks para descrever o conteúdo dos dados geográficos e finalmente ele identifica dados sobrepostos disponíveis em diferentes sítios da Web.

3.1. Texto âncora e texto âncora estendido

Normalmente os atributos dos arquivos (nome, data da criação e última modificação, tamanho e tipo) não oferecem descrições detalhadas do conteúdo do arquivo. A falta de informações adicionais torna a indexação destes arquivos uma tarefa difícil e algumas vezes os resultados obtidos não são os desejáveis.

Um sítio da Web pode ser composto por multimídias tais como sons, imagens, textos, arquivos e pelas conexões a outros sítios ou páginas, os hyperlinks. A estrutura criada por estas conexões está sendo pesquisada e utilizada para melhorar a eficiência dos rastejadores [Cho, Garcia-Molina and Page 1998] e o processo de classificação de páginas utilizada pelos mecanismos de busca, para descobrir comunidades Web e para organizar os resultados da pesquisa em hubs e autoridades. Um hyperlink contém a URL para a página que ele referencia e um texto âncora que descreve a ligação. O texto âncora pode oferecer excelentes descrições das páginas que ele referencia. Estes textos âncoras podem ser úteis para descrever e auxiliar na recuperação de páginas não indexadas, que contêm elementos como imagens, arquivos de banco de dados e dados geográficos, por mecanismos de busca tradicionais.

A idéia de utilizar texto âncora foi inicialmente implementada na World Wide Web Worm [Mcbryan 1994] especialmente porque ela auxilia na busca de informações não textuais. O texto âncora permite conectar palavras (e contexto) a um conteúdo específico (por exemplo, clique aqui para obter o mapa da cidade de São José dos Campos na escala 1:20.000).

O mecanismo desenvolvido utiliza o conceito de texto âncora para auxiliar na descrição do contexto e nos resultados da busca. Para melhorar os resultados obtidos utiliza também o texto âncora estendido. Neste caso, além do texto do link, as palavras e frases próximas dos links são consideradas para classificar os dados. A figura 1 ilustra os conceitos de texto âncora e texto âncora estendidos.

Devido ao tamanho dos arquivos de dados geográficos, normalmente estes arquivos estão disponíveis na Web em formatos compactados na forma de arquivos zip, arj, rar e outros. Estes formatos não estão no foco de nosso mecanismo. Neste caso, o texto âncora tem outra importante função: ajudar os rastejadores localizar arquivos compactados de dados geográficos a partir da análise do contexto das páginas.

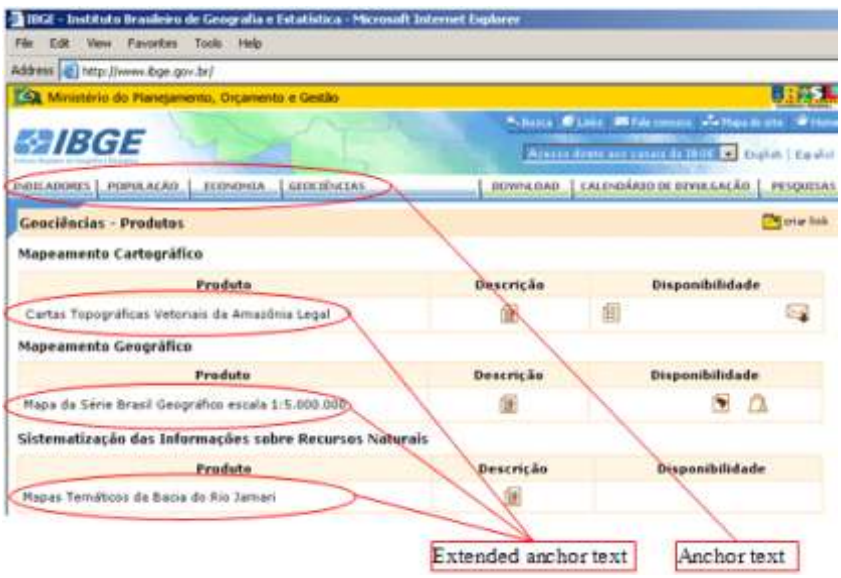

Figura 1. Texto âncora e texto âncora estendido.

3.2. Processamento distribuído

Para suportar todas as requisições de busca e consulta, nosso protótipo está baseado em processamento distribuído com um servidor de aplicações centralizado com Web services. Esse servidor é responsável pelo gerenciamento das requisições dos clientes. Atualmente os principais mecanismos de busca têm dezenas de bilhões de páginas indexadas em seu banco de dados, dessa forma, o processo de extração, análise e indexação dos sítios Web é lento. Outro problema é o tempo de revisita que acontece entre 30 e 60 dias, tornando o banco de dados desatualizado. Utilizar processamento distribuído com usuários colaboradores pode tornar estes processos mais eficientes, rápidos e exigir menores investimentos em infra-estrutura física.

3.3. Usuários colaboradores

Usuários colaboradores são clientes que ajudam no processamento quando seus computadores estão ociosos. Eles contribuem com o mecanismo desenvolvido rastejando a Web e executando a análise sintática para encontrar evidências de dados geográficos. Para se tornar um colaborador, o usuário precisa instalar o módulo que gerencia as tarefas relacionadas ao mecanismo de busca.

Dentre as vantagens de trabalhar com usuários colaboradores destacam-se a redução de investimentos para manter o projeto, uma vez que o processamento distribuído torna desnecessária a utilização de servidores poderosos para executar as funções de busca e análise e a capacidade de crescimento sustentável à medida que novos usuários colaboradores aderem ao projeto.

3.4 Outras características

Outra importante característica do mecanismo proposto é que seus rastejadores respeitam o código de ética dos rastejadores. Para evitar que um sítio específico seja indexado, são observadas algumas meta-tags tais como <meta name = "robots" content = "noindex, nofollow">. Esta meta-tag é incluída no cabeçalho das páginas e o valor robots pode ser alterado pelo nome de um mecanismo específico (por exemplo google ou yahoo). O valor noindex estabelece para o robô que a página não pode ser indexada. O valor nofollow determina que os links eventualmente existentes na página não podem ser seguidos. Qualquer arranjo dos valores index/noindex, follow/nofollow é permitido.

4. Funcionamento do Sistema

O mecanismo desenvolvido é implementado em C# e o servidor de banco de dados utiliza o SQL Server 2000. Existem duas arquiteturas principais: Servidores e Clientes. O papel básico do primeiro é gerenciar a distribuição de URLs, receber, organizar e armazenar os dados e arquivos capturados, e disponibilizar uma interface para o usuário executar consultas na Web e visualizar os resultados obtidos. Os clientes, executados nos computadores dos usuários colaboradores, desempenham os seguintes papéis: requisitar uma lista de URLs para serem visitadas, rastejar os sítios indicados, capturar as páginas, analisar o conteúdos das páginas, extrair os dados de interesse e enviar o conteúdo encontrado para o servidor. A figura 2 demonstra o funcionamento do sistema.

O tráfego de dados entre os clientes e o servidor é executado utilizando XML (eXtended Markup Language) e Web services para possibilitar a utilização da aplicação por computadores protegidos por firewall e proxy.

O processo para descobrir dados geográficos é iniciado quando o computador de um usuário colaborador está ocioso. O cliente solicita uma lista de URLs para serem visitadas. O servidor de Web service (WS) envia um lista ordenada de URLs para o cliente. O cliente utiliza o rastejador para encontrar os sítios indicados e recuperar o conteúdo HTML das páginas visitadas. Então, seu conteúdo é analisado visando

encontrar dados geográficos e extraindo os dados desejados (URLs, endereços dos dados geográficos, palavras-chave, texto âncora e texto âncora estendido), e enviando esses dados para o WS que armazena os dados organizados no servidor de banco de dados (BD). Antes de serem armazenadas as URLs relativas são convertidas para URLs absolutas.

Novas URLs são ordenadas para serem visitadas posteriormente. A função de indexação é executada pelo indexador. O indexador executa várias funções: faz a leitura do repositório e o analisa, analisa também todos os links das páginas e armazena informações importantes sobre eles em arquivos específicos. Estes arquivos contêm informação suficiente para determinar para onde cada link aponta e o texto do link.

O servidor BD está conectado ao servidor de downloads (SD) por uma rede interna. O SD monitora os caminhos de arquivos incluídos no servidor de banco de dados e inicia o processo de download para cada inclusão. O SD busca o arquivo em seu local original, o captura e o armazena em um diretório. Então o arquivo é compactado para otimizar o espaço de armazenamento. Para a compactação e descompactação dos arquivos é utilizada a classe GZipStream disponível no .NET framework 2.0.

Sobre os arquivos armazenados é realizada a análise específica para os dados geográficos, buscando a projeção e o nome do local nas colunas dos arquivos DBF. Essa informação é ordenada e armazenada em um repositório do servidor BD e posteriormente é utilizada no processo de busca. Informações adicionais relevantes presentes nas colunas também são armazenadas.

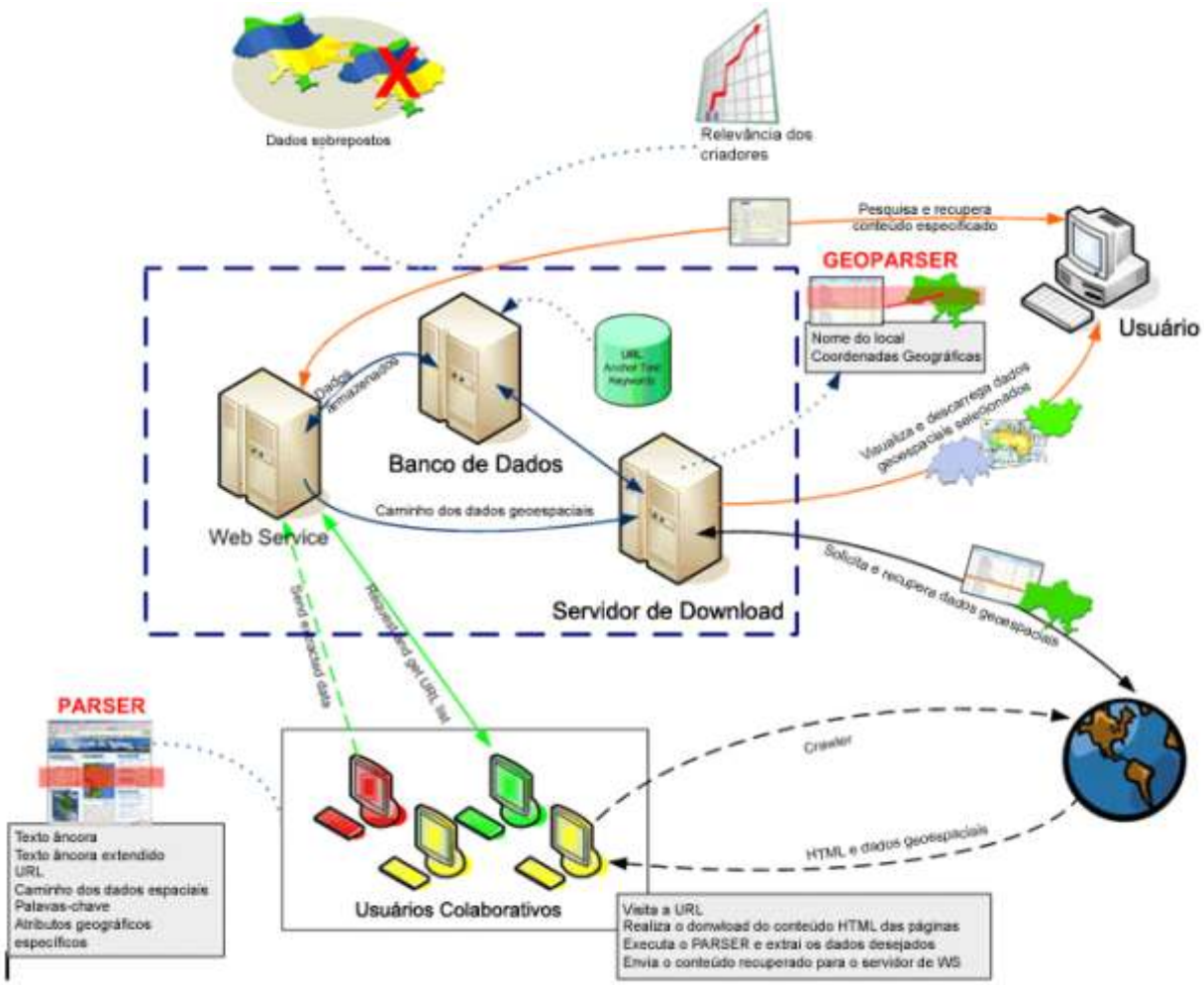

Figura 2. Funcionamento do mecanismo de busca.

4.1. Rastejador

O rastejador é responsável por visitar os endereços armazenados no servidor BD. A primeira tarefa executada pelo rastejador é a verificação do arquivo robots.txt no diretório raiz do servidor que hospeda o sítio visitado para certificar-se que não exista nenhuma regra que impeça a indexação das páginas. Informações detalhadas sobre o arquivo robots estão disponíveis em http://www.robotstxt.org/wc/norobots.html.

Após executar esta tarefa e certificar-se que a página pode ser indexada, o rastejador faz uma cópia de seu conteúdo por meio de um protocolo http. O rastejador utiliza classes que retornam um stream de bytes dos dados recebidos. Esses dados são convertidos para caracteres ASCII e o conteúdo da página é reconstruído de forma que o analisador possa executar o processo de análise e extração das informações.

O rastejador é uma aplicação complexa pois demanda a interação com milhares de servidores Web que extrapolam o controle do sistema [Brin and Page 1998]. Para visitar milhões de sítios da Web, o mecanismo desenvolvido tem um sistema distribuído de rastejadores que roda nos computadores dos usuários colaboradores. Atualmente cada rastejador visita 60.000 páginas por dia. Este é um excelente número uma vez que nosso mecanismo pode envolver centenas de usuários colaboradores aumentando consideravelmente a cobertura das páginas disponíveis na Web.

4.2. Analisador

O analisador é responsável pela extração de todo o conteúdo das páginas. Nestas páginas estão os endereços que serão armazenados na base de dados e posteriormente visitados pelos rastejadores. Assim que os rastejadores executam o download das páginas, todos caracteres são convertidos para minúsculos. O analisador verifica todo o conteúdo HTML para extrair algumas informações relevantes para a indexação da página. Entre este conteúdo estão as meta-tags: <title></title> que descreve o título da página, <META NAME="Description" Content=""> que descreve o conteúdo da página, <META NAME="Keywords" Content=""> que armazena palavras-chave relacionadas ao conteúdo da página, <a href=""> que indica links para outras páginas e arquivos de dados geográficos.

O analisador remove todas as tags HTML, scripts e outras marcas mantendo o texto puro. As palavras extraídas são armazenadas em uma tabela de palavras no servidor BD. Para cada palavra um código hash é gerado para melhorar a velocidade no processo de busca.

4.3. Armazenador de endereço

O armazenador de endereço é responsável pelo armazenamento, no servidor BD, dos endereços extraídos pelo analisador e pela verificação do correto armazenamento dos endereços. Os endereços são utilizados pelo servidor WS para distribuir as URLs que serão visitadas e analisadas pelos usuários colaboradores.

4.4. Busca

O processo de busca é focado na qualidade dos resultados obtidos. A interface com o usuário é amigável e roda em navegadores tradicionais.

Para a execução da busca, o servidor WS verifica no servidor BD as palavras-chave relacionadas com as informadas pelo usuário. Quando o servidor WS encontra resultados que satisfaçam a consulta, ele retorna para o usuário uma lista de metadados que atendam o conteúdo de sua busca. A classificação dos arquivos na lista é calculada considerando a proximidade dos termos de consulta com os termos encontrados no servidor BD e pela relevância do produtor.

Para cada item presente na lista de arquivos, são fornecidas informações adicionais tais como

relevância do produtor, quantidade de dados geográficos disponíveis no sítio deste produtor, quantidade de downloads solicitados do arquivo. Estas informações auxiliam o usuário a conhecer melhor os dados geográficos antes de iniciar o download.

Os passos para a busca são:

1. *O usuário acessa a interface do mecanismo e digita as palavras-chave para a busca.*
2. *O servidor WS analisa as palavras.*
3. *O servidor BD é percorrido até que exista um arquivo de dados geográficos que satisfaça as palavras-chave. Este passo é repetido até percorrer todo o servidor.*
4. *A classificação dos arquivos selecionados é calculada e os arquivos são ordenados.*
5. *O servidor WS envia uma lista de arquivos classificados e informações adicionais para o usuário.*

O download dos arquivos pode ser executado através do servidor de downloads ou a partir de seu local de origem.

4.4. Busca

A qualidade dos resultados da busca é a principal medida de um mecanismo de busca. O GeoDiscover produz excelentes resultados ao manipular dados geográficos. A partir de somente uma URL armazenada em seu banco de dados, O GeoDiscover recupera a página, captura novas URLs e executa visitas em um processo infinito e ininterrupto. Em testes preliminares, o GeoDiscover processou 60.000 páginas por dia utilizando um computador com 2.0 Ghz de processamento, 512 Mb de memória RAM e uma conexão de internet de 128 Mbps.

5. Conclusões

Este mecanismo foi projetado para tornar melhorar o acesso e recuperação de arquivos de dados geográficos disponíveis na Web. Ele foi implementado em um ambiente distribuído com usuários colaboradores. Os usuários colaboradores são muito interessantes pois auxiliam nos processos de recuperação e análise das páginas aumentando significativamente a cobertura da Web. Este cenário apresenta algumas vantagens: redução de investimentos para manter o mecanismo, uma vez que se torna desnecessário servidores poderosos e grande capacidade de expansão do mecanismo à medida que novos usuários façam parte do projeto.

Para melhorar os resultados das buscas foram utilizadas técnicas para análise e recuperação de dados geográficos, dados sobrepostos, descrição de relevância dos produtores de dados, texto âncora e texto âncora estendido. O mecanismo apresentado possui funções para coletar, indexar e executar buscas em dados geográficos.

Algumas funções estão sendo desenvolvidas para completar o projeto, dentre elas a utilização de gazetteers para comparar nomes de lugares; a ampliação dos rastejadores a fim de reconhecer outros formatos GIS tais como arquivos spr e geotiff, a inclusão de ontologias para melhorar os resultados retornados à consulta dos usuários, a implementação de um sistema amigável de pré-visualização dos arquivos de dados geográficos através do navegador.

Abstract

The fast development of the internet and the growth of digital contents available led to the development of search engines that facilitate the recovery of information in the Web. However, these engines have limitations mainly when recovering specialized contents. Geospatial data are created by local governments, companies and people, but these data aren't available in a systematized way. In this context,

this paper presents a specialized search engine to access and recover geospatial data in the Web, focusing on its main characteristics, architecture, technologies and performance.

Keywords:
niche search engine - geospatial data - web.

6. Referências Bibliográficas

Berners-Lee, T., J. Hendler and O. Lassila (2001). "The Semantic Web." Scientific American May.

Brin, S. and L. Page (1998). The anatomy of a large-scale hypertextual Web search engine. Seventh International World Wide Web Conference, Brisbane, Australia.

Cho, J., H. Garcia-Molina and L. Page (1998). Efficient Crawling Through URL Ordering. Seventh International Web Conference (WWW 98). Brisbane, Australia.

Egenhofer, M. (2002). Toward the semantic geospatial web. 10th ACM international symposium on Advances in geographic information systems table of contents, McLean, Virginia, USA.

Frew, J. (1998). "The Alexandria Digital Library Architecture." International Journal on Digital Libraries 2(4): 259-268.

Gerhart, A. (2002). "Understanding and Building Google PageRank." from http://www.searchengineguide.com/orbidex/2002/0207_orb1.html

Giles, C. L., K. Bollacker and S. Lawrence (1998). CiteSeer: An automatic citation indexing system. Digital Libraries 98 - The Third ACM Conference on Digital Libraries, Pittsburgh, PA, ACM Press.

Glover, E. J. T., K.; Lawrence, S.; Pennock, D.; Flake, G. W (2002). Using Web Structure for Classifying and Describing Web Pages. WWW2002, Honolulu, Hawaii, USA.

Gruber, T. R. (1995). "Toward Principles for the Design of Ontologies Used for Knowledge Sharing." Int. Journal of Human-Computer Studies 43: 907-928.

Guarino, N. and P. Giaretta (1995). Ontologies and Knowledge Bases: Towards a Terminological Clarification. Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing 1995, Amsterdam, IOS Press.

Howe, W. (1996). "When did the Internet start? A brief capsule history." Disponível em: http://intranet.canacad.ac.jp/instruction/internet_skills/brief_history/!history.html Acesso em: 20/10/2005.

Lawrence, S. and C. L. Giles (1998). "Searching the World Wide Web." Science 280(5360): 98-100.

Lawrence, S. and C. L. Giles (1999). Text and Image Metasearch on the Web. International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA 99.

Masolo, C., A. Gangemi, N. Guarino, et al. (2002). The WonderWeb Library of Foundational Ontologies. Padova, LADSEB-Cnr: 36.

Mcbryan, O. A. (1994). GENVL and WWWW: Tools for Taming the Web. First International Conference on the World Wide Web, Geneva, CERN.

Onsrud, H., G. Camara, J. Campbell, et al. (2004). Public Commons of Geographic Data: Research and Development Challenges. III International Conference on Geographical Information Science (GIScience 2004), Washington, Springer.

Page, L., S. Brin, R. Motwani, et al. (1999). "The PageRank Citation Ranking: Bringing Order to the Web." from <http://dbpubs.stanford.edu/pub/1999-66>.

Wiederhold, G. (1994). Interoperation, Mediation and Ontologies. International Symposium on Fifth Generation Computer Systems (FGCS94), Tokyo, Japan, ICOT.

Sobre os autores:

Fernando Renier Gibotti
é Bacharel em Ciência da Computação, Especialista em Geoprocessamento, Mestre em Engenharia Urbana e aluno de Doutorado em Computação Aplicada no Instituto Nacional de Pesquisas Espaciais (INPE). Atualmente é coordenador e professor do curso de Sistemas de Informação da UniFAIMI em Mirassol - SP. E-mail: gibotti@dpi.inpe.br

Gilberto Câmara
é Graduado em Engenharia Eletrônica pelo ITA, tem Mestrado e Doutorado em Computação pelo INPE. Trabalha desde 1980 no INPE e lidera a equipe de P&D em Geoprocessamento do INPE, responsável pelo desenvolvimento dos softwares SGI e SITIM (1981-1993), SPRING (1991-presente), e TerraLib (2002-presente). Foi chefe da Divisão de Processamento de Imagens (1991-1996), e coordenador-geral da área de Observação da Terra (OBT) de outubro/2001 a novembro/2005. Atualmente é diretor do INPE para o período de dezembro/2005 a dezembro/2009. E-mail: gilberto@dpi.inpe.br

Renato Almeida Nogueira
é bacharel em Sistemas de Informação pela UniFAIMI. Atualmente trabalha com desenvolvimento de sistemas para negócios. E-mail: renato@faimi.edu.br