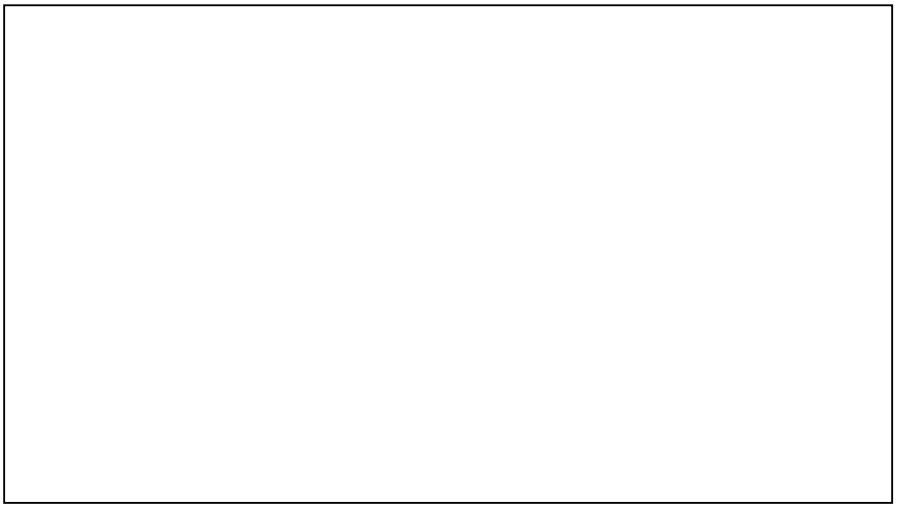
MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

GEODISCOVER - MECANISMO DE BUSCA ESPECIALIZADO EM DADOS GEOGRÁFICOS

Fernando Renier Gibotti da Silva

Tese de Doutorado em Computação Aplicada, orientada pelo Dr. Gilberto Câmara.

INPE
São José dos Campos
2006



FOLHA DE APROVAÇÃO

“Descobre-se que se leva muito tempo
para ser a pessoa que se quer ser
Que o tempo é curto...”
O Menestrel – Shakespeare

A meus pais,
Maria Ermelinda Gibotti da Silva e
José Luis Borges da Silva

As minhas principais motivações,
Denise e Fernanda

AGRADECIMENTOS

Agradeço ao Dr. Gilberto Câmara pelo apoio incondicional, por me ajudar a encontrar um caminho onde eu pudesse contribuir e pelos momentos de discussão que sempre ampliaram meus horizontes. A ele minha eterna gratidão e admiração.

Aos professores e pesquisadores da DPI, pelo espírito de equipe e pelos conhecimentos que me transmitiram, muito obrigado. Especialmente ao Dr. Antonio Miguel e Dr. Gerard Banon pelo apoio e discussões oportunas.

Ao professor Frederico Fonseca pelo acolhimento na University of Maine (onde tudo começou), pela significativa contribuição na estruturação desta tese e pelas discussões mediadas com o Dr. Lee Giles na Penn State University.

Minha gratidão à direção e colegas da UniFAIMI, pelo apoio decisivo e pela confiança no meu trabalho. Meus sinceros agradecimentos ao Renato Almeida pela contribuição na implementação e ao Thiago Simonato pelo companheirismo e suporte constantes.

Ao André Baldo, Estevão Medeiros, Heimdall Bergamini, Lucas Bertoni e Paulo Fiaschi pela ajuda nos testes da base de dados e interface. À Moema pela atenciosa leitura do texto.

Ao INPE pela oportunidade de aperfeiçoamento acadêmico.

Agradeço aos amigos do INPE Gilberto Ribeiro, Marcelino, Tiago Carneiro e Claudia Santos pelo convívio, auxílio e pelas palavras de motivação nos momentos mais delicados de minha vida nos últimos cinco anos.

Ao meu pai e minha mãe pelo suporte, carinho e incentivo. Ao Alexander, meu irmão, que sempre vibrou com minhas conquistas. À Denise pelo carinho e compreensão.

RESUMO

Dados geográficos estão sendo amplamente utilizados para subsidiar análises e tomadas de decisões em diferentes domínios do conhecimento. O desenvolvimento acelerado da Internet e o aumento de conteúdos digitais disponíveis propiciaram 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 são produzidos por empresas estatais, privadas e por pessoas, mas esses dados não estão disponibilizados de forma sistematizada e não são encontrados na *Web* por mecanismos convencionais de forma eficiente, pois tais mecanismos não estão preparados para evidenciar este conteúdo. Nesse contexto, esta tese apresenta um mecanismo de busca especializado em descoberta, indexação e classificação de arquivos geográficos disponíveis na *Web*. Para tanto, foram estabelecidos métodos para descobrir arquivos geográficos disponíveis no ciberespaço por meio da análise combinada do contexto das páginas *Web* com o conteúdo dos arquivos geográficos. Para a consulta e recuperação dos arquivos indexados foi desenvolvida uma interface amigável. Para avaliar o desempenho do mecanismo implementado em termos de qualidade e abrangência, foram executados testes utilizando os métodos de *abrangência* e *precisão*.

GEODISCOVER – A NICHE SEARCH ENGINE TO DISCOVER GEOSPATIAL DATA IN THE WEB

ABSTRACT

Geospatial data are being broadly used to provide analysis and decisions in different domains of the knowledge. 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 and they aren't found by conventional search engines in an efficient way therefore such engines aren't prepared to evidence this content. In this context, this thesis presents a niche search engine specialized in discover, indexing and classification of geospatial data available in the web. We develop methods to discover geospatial data available in the web analyzing together the context of web pages and the content of the geospatial data. A friendly-user interface was developed to query and recover the indexed files. In order to evaluate the search engine performance we used the measures of information retrieval – *Recall* and *Precision*.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ASCII - American Standard Code for Information Interchange

BD – Banco de dados

DAML - DARPA Agent Markup Language

GeoTIFF – TIFF for georeferenced raster imagery

GML – Geographic Markup Language

HTML – Hyper Text Markup Language

HTTP – Hyper Text Transfer Protocol

OGC – Open Geospatial Consortium

OIL - Ontology Inference Layer

OWL –Web Ontology Language

PDF – Portable Document Format

WS – Servidor de Web Service

TI – Tecnologia da Informação

TIFF – Tagged Image File Format

URL – Uniform Resource Locator

CAPÍTULO 1

INTRODUÇÃO

Desde sua origem como projeto de pesquisa no final da década de 60, a Internet cresceu rapidamente e hoje é o principal veículo de disseminação de informação. Como explica (Lessing, 1999), a natureza aberta dos protocolos da Internet (TCP/IP, SMTP, HHTP) permitiu que a criação de aplicativos e a disseminação ampla de dados, sem restrições de propriedade intelectual. A partir da base tecnológica da Internet, criou-se a rede mundial de informações (*World Wide Web* ou apenas *Web*). A *Web* é hoje uma rede composta por bilhões de páginas multimídia interligadas, desenvolvidas de forma independente por milhões de pessoas. Este crescimento explosivo e não-estruturado tem seu preço, quando se procura recuperar informações na rede. A mesma natureza aberta dos protocolos que permitiu o rápido crescimento da Web, dificulta hoje a busca de informações associadas a conteúdo.

Para melhorar a procura e a análise de informações na *web,* foram desenvolvidas ferramentas de busca como Google (www.google.com), Altavista (www.altavista.com) e Yahoo (www.yahoo.com) (Glover, 2002). Os mecanismos de busca mais utilizados atualmente, embora poderosos, possuem limitações no que diz respeito à informação especializada, principalmente na quantidade de arquivos indexados e na qualidade nos resultados da pesquisa. Para superar essas limitações, há atualmente um grande esforço para organizar a informação semântica associada às páginas da *web.* A partir da visão apresentada por Berners-Lee em seu paper *The Semantic Web* (Berners-Lee *et al.*, 2001), está havendo um grande esforço para propor uma estrutura para as páginas web que permita sua recuperação eficiente, com base em busca semântica. Estas estruturas são baseadas na idéia de descrição compartilhada do conteúdo das páginas, através de léxicos padronizados, também chamados de *ontologias.*

A pesquisa em Ontologias começou na comunidade de Inteligência Artificial (Gruber, 1995), (Guarino; Giaretta, 1995), (Wiederhold, 1994), como forma de representação de conhecimento. A idéia geral é que a descrição formal do significado dos termos usados em um sistema computacional permite o seu compartilhamento. A principal limitação do uso de

Ontologias é a sua natureza estática. Sabemos que parte do conhecimento está relacionado à dinâmica das ações humanas. Deste modo, compartilhar apenas as ontologias (léxico) não permite uma representação completa do conhecimento. No entanto, dada a natureza estática das páginas da *Web* e dada a falta de protocolos para compartilhar semântica, a idéia de Ontologias surgiu como uma forma de buscar estruturar o caos. Assim, a comunidade de Tecnologia de Informação está trabalhando em propostas para organizar a informação visando à interoperabilidade e à fácil recuperação de informações. Essas propostas incluem linguagens tais como OWL (Masolo *et al.*, 2002), GML (OGC, 1999) e DAML-OIL (Bechhofer *et al.*, 2000).

Na área de geoinformação, os benefícios potenciais do compartilhamento de dados são grandes (Egenhofer, 2002) (Fonseca; Engenhofer, 1999). Sabe-se que a coleta e conversão de dados representam mais de 50% dos custos de um projeto de geoinformação (Egenhofer; Frank, 1990). Assim, a comunidade de geoinformação trabalha na criação de uma *Geospatial Semantic Web* (Egenhofer, 2002). Como exemplo, o *Open Geospatial Consortium* propôs a linguagem GML para compartilhar dados espaciais. Entretanto, a utilização de linguagens tais como OWL ou GML requer que os produtores de dados produzam e organizem ontologias. Trata-se de uma tarefa que pode estar além das habilidades e possibilidades de muitas organizações. Além disso, já existe uma enorme quantidade de dados espaciais disponíveis na *web*. Não é razoável esperar que todos esses dados sejam convertidos para os padrões da *Semantic Web* para poder utilizá-los.

Uma alternativa para aproveitar os dados já existentes é fornecer mecanismos de busca especializados. Estes mecanismos sabem que algumas comunidades produzem documentos estruturados e que, a partir de hipóteses razoáveis sobre sua estrutura, muitas informações semânticas podem ser recuperadas diretamente. Os exemplos mais recentes são os mecanismos para recuperar e indexar artigos científicos, tais como Google Scholar e Citeseer (Giles *et al.*, 1998). Essas ferramentas reconhecem a estrutura bem definida de um artigo científico, com título, lista de autores, resumo, texto e referências. A partir desta estrutura, montam bancos de referências bibliográficas e indexam citações. Desta forma, os mecanismos especializados fornecem informação útil e o único esforço exigido dos autores é a postagem de seus artigos em uma página *web*.

E quanto aos 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 esses dados e os usuários. No entanto, 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. Como acontece em artigos científicos, em que os arquivos PDF são predominantes, os dados geoespaciais usualmente são distribuídos em formatos conhecidos, como *shapefiles* e GeoTIFF. Estes formatos incluem informação descritiva sobre os dados de forma simplificada. Estas características de dados geográficos tornam possível adaptar a idéia de mecanismos como o *Google* Scholar para dados geográficos.

Com base na motivação acima, este trabalho descreve uma ferramenta de busca especializada em dados geoespaciais, que considera a natureza semi-estruturada destes dados. A idéia é permitir o compartilhamento dos dados geográficos sem a necessidade de trabalho adicional com anotações semânticas. O mecanismo de busca proposto é especializado em descoberta, indexação e classificação de arquivos geográficos disponíveis na *web*. Para tanto, apresenta métodos para descobrir arquivos geográficos disponíveis no ciberespaço, por meio da análise combinada do contexto das páginas *web* juntamente com o conteúdo dos arquivos geográficos. Também propõe uma interface amigável para que usuários leigos possam recuperar arquivos de interesse de forma rápida e precisa.

Para efetuar o processo de descoberta, classificação, pesquisa e visualização de arquivos geográficos, implementamos uma arquitetura distribuída, baseada em usuários colaboradores, com rastejadores[1] e métodos para análise de contexto geográfico, para extração de metadados, classificação dos produtores de dados e visualização de resultados de consultas.

[1] Neste trabalho foram utilizados os termos “rastejadores” como tradução de *Web Crawler* e “análise de contexto geográfico” para *parser*.

1.1 Questão Científica, Hipótese de Trabalho e Resultados

Com base no contexto acima exposto, a questão científica da tese é: *como encontrar, indexar e disponibilizar arquivos geográficos no ciberespaço utilizando a estrutura atual da web?*

A hipótese de trabalho adotada faz analogia ao *Google Scholar*. Consideramos que todo arquivo geográfico tem um mínimo de informação a respeito de seu conteúdo. Assim, pode-se desenvolver rastejadores especializados para recuperação de arquivos geográficos[2] na *web*. Como base nesta hipótese, desenvolvemos uma ferramenta capaz de descobri-los, indexá-los e disponibilizá-los, com os seguintes passos:

- Elaboração de metodologia para a análise combinada da estrutura do arquivo geográfico com o conteúdo de páginas *web* objetivando extrair informações semânticas que permitam identificar arquivos geográficos;
- Criação de um repositório de palavras geo-interessantes que denotem indícios de arquivos geográficos, para ser utilizado na análise sintática executada das páginas *web*;
- Construção de rastejadores especializados distribuídos em usuários colaboradores utilizados para descobrir arquivos geográficos no ciberespaço;
- Definição de padrões de interfaces amigáveis que permitam consultas por navegadores e apresentem resultados de forma clara e objetiva;
- Utilização de métricas de qualidade para avaliar a eficiência do sistema de recuperação de informação.

A partir da metodologia proposta, foram obtidos os seguintes resultados:

- Desenvolvimento de um protótipo computacional capaz de descobrir, analisar e classificar arquivos geográficos disponíveis na *web*;

[2] Com vistas às três dimensões básicas para informação geográfica Fonseca, F.; Sheth, A. **The geospatial semantic web**. University Consortium for Geographic Information Science, 2003. , o GeoDiscover procura identificar arquivos da dimensão *profissional*: "informação geográfica estruturada armazenada em banco de dados geográficos que estão indexados ou descritos em páginas *web*".

- Aplicação da metodologia de análise combinada da estrutura do arquivo geográfico com o contexto das páginas *web* que apontam para o arquivo, para retornar respostas mais precisas às consultas efetuadas pelos usuários do *GeoDiscover*.

1.2 Contribuição da Tese

Os mecanismos de busca utilizados atualmente cobrem parte significativa da *web* e são muito utilizados para consultas genéricas. Recentemente vários projetos estão relacionados à criação de mecanismos de busca por nichos. Dentre eles, alguns se especializaram na busca do conteúdo geográfico presente nas páginas *web* para direcionar os resultados de consultas a partir de uma referência geográfica.

Esta tese apresenta uma proposta para encontrar, indexar e disponibilizar para os usuários de geoinformação, arquivos geográficos disponíveis na *web* utilizando:

Este trabalho contribui para superar esta limitação utilizando:

- Metodologia para a descoberta e captura de arquivos geográficos a partir da análise do contexto da página em busca de indícios de conteúdo geográfico com a utilização de rastejadores especializados em arquivos *shape* e *zip* e;

- Rastejadores distribuídos em usuários colaboradores para ampliar a cobertura da *web*, diminuir o tempo utilizado para visita e possibilitar crescimento sustentável do mecanismo;

- Técnicas de indexação e classificação dos arquivos utilizando análise combinada do conteúdo do arquivo geográfico com o contexto da página que aponta para o arquivo através da análise do texto âncora e texto âncora estendido do *hyperlink*;

- Técnicas para a classificação dos resultados a partir da qualificação dos produtores de dados.

- Interface gráfica amigável que apresenta atributos e pré-visualização gráfica dos arquivos resultantes do processo de consulta.

Esta tese é inédita e diferencia-se de outros trabalhos correlatos desenvolvidos, pois dá ênfase à busca específica de arquivos geográficos disponíveis na *web*, enquanto que os demais utilizam o conteúdo geográfico presente em páginas *web* para permitir que usuários façam buscas baseadas em contexto geográfico.

1.3 Organização da Tese

Esta tese concentrou-se em técnicas para a descoberta, indexação e busca de arquivos geográficos disponíveis na *web*. A Figura 1-1 ilustra os principais tópicos abordados neste trabalho, o qual está organizado em sete capítulos.

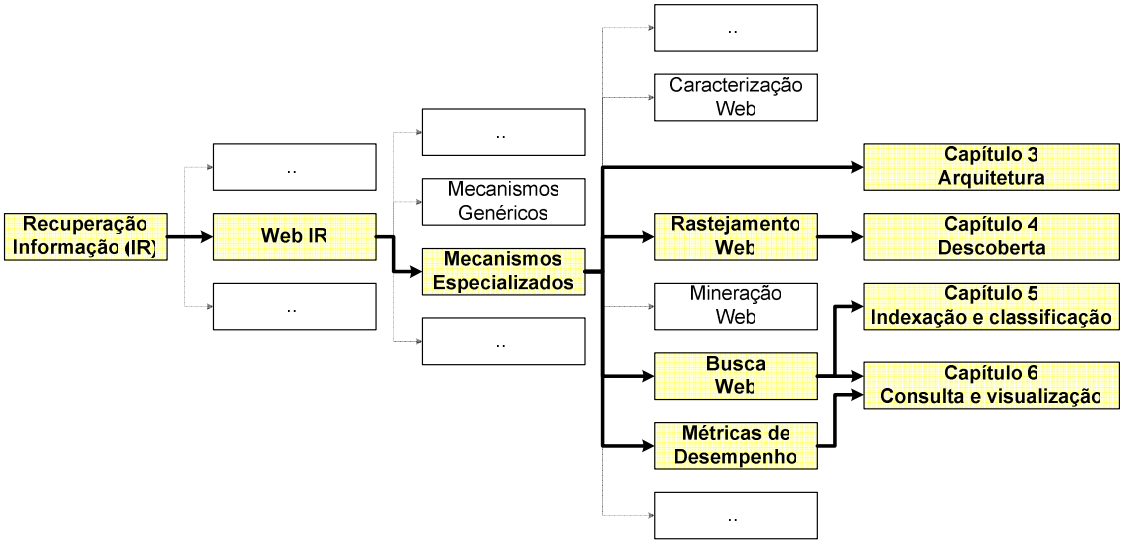


Figura 1-1 – Tópicos principais abordados na tese.

O capítulo 2 apresenta a fundamentação teórica deste trabalho, abordando a recuperação de informações, os mecanismos de busca para recuperação de informação na web, as métricas de qualidade e recuperação de informações, as peculiaridades na busca de arquivos geográficos e os mecanismos de busca especializados em informações geográficas.

O capítulo 3 exibe a anatomia de um mecanismo de busca especializado em arquivos geográficos com duas visões distintas: a visão conceitual, que focaliza a organização dos processos de descoberta, classificação, indexação e recuperação de dados, e a visão estrutural, que discorre sobre a estrutura física dos servidores e do usuário colaborador. Discute a metodologia utilizada para a descoberta de arquivos geográficos disponíveis na

web e apresenta o rastejador específico para arquivos *shape*, a análise sintática das páginas *web* em busca de indícios de arquivos geográficos e a criação dos repositórios de palavras geo-interessantes e de termos descritores de lugares.

O capítulo 4 volta-se para a implementação e avaliação do *GeoDiscover*. Apresenta as técnicas para consulta e visualização dos arquivos que satisfaçam à busca e a ordenação dos resultados para a visualização, bem como o desempenho do protótipo implementado.

O capítulo 5 apresenta algumas considerações finais da tese, as limitações e trabalhos futuros.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica para esta tese. Serão apresentados conceitos e definições sobre recuperação de informações, recuperação de informações na web com a utilização de mecanismos de busca detalhando seus componentes básicos, medidas de qualidade em recuperação de informações, peculiaridades de arquivos geográficos e mecanismos de busca especializados em contexto geográfico.

2.1 Recuperação de Informação

O termo original "*information retrieval*" foi cunhado pelo pesquisador americano **Calvin Northrup Mooers** em 1952 (Jones; Willett, 1997). Recuperação de informação é o nome dado ao processo ou método pelo qual um usuário é capaz de converter sua necessidade por informação em um conjunto de documentos armazenados que contenham informações de interesse.

Um sistema de recuperação de informações é responsável pelo armazenamento e gerenciamento de informações em diferentes tipos de documentos e pode auxiliar os usuários a encontrarem informações de interesse. O objetivo do sistema é informar a existência e localização de documentos que possam conter a informação necessária e não necessariamente recuperar a informação. Conforme narrado por Baeza-Yates (2004):

> *"... the IR system must somehow 'interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query. This 'interpretation' of a document content involves extracting syntactic and semantic information from the document text..."*

Existem três processos básicos em um sistema de recuperação de informação: a representação do conteúdo dos documentos, a representação das necessidades de informação do usuário e a comparação entre as duas representações (Croft, 1993). A Figura 2-1 ilustra os processos de um sistema de recuperação de informação, representado pelos retângulos com bordas arredondadas.

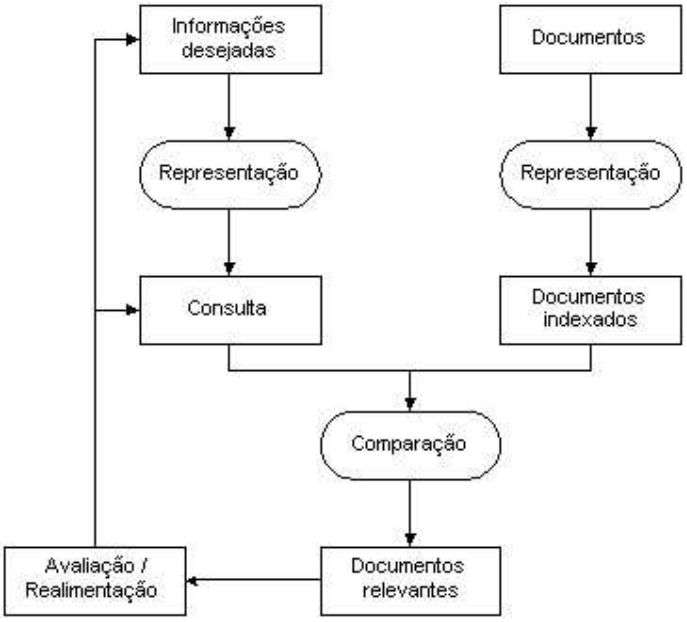


Figura 2-1 – Processo de recuperação de informação.
Adaptado de (Croft, 1993)

A *representação de documentos* é utilizada para construir o sistema. Produz um índice organizado, que pode incluir o documento completo e seus metadados (como o título, palavras-chave e o resumo). O processo de *representação das informações desejadas* está associada a uma solicitação do usuário. Esta solicitação resulta em uma expressão de consulta. A *comparação* resulta em uma lista de documentos ordenados pela sua relevância, na qual o usuário pode navegar e encontrar a informação de que necessita. A ordenação pela relevância é crítica, pois pode reduzir drasticamente o tempo que o usuário irá despender para encontrar as informações de interesse.

Quando uma consulta é solicitada a um sistema de RI, os documentos relacionados podem ser divididos conceitualmente em duas categorias: uma de documentos relevantes e outra de documentos não relevantes. Dada uma expressão de busca, o sistema computa uma medida de relevância (denominada *Retrieval Status Value* ou RSV) para cada item da coleção de documentos (Raghavan *et al.*, 1989). O RSV é um indicador do grau de similaridade entre um documento e uma consulta. Permite ordenar a coleção de documentos para que o sistema possa decidir sobre quais itens devem ser recuperados.

Um dos maiores usos de recuperação de informação está associado à Internet. Este caso *web* possui características específicas que são tratadas pelos mecanismos de busca. Eles utilizam um rastejador para vasculhar o ciberespaço e encontrar documentos que possam ser indexados. Oferecem ainda uma interface acessível em um *browser* para que o usuário possa efetuar suas consultas. A Figura 2-2 apresenta os componentes para a recuperação de informações na *web*.

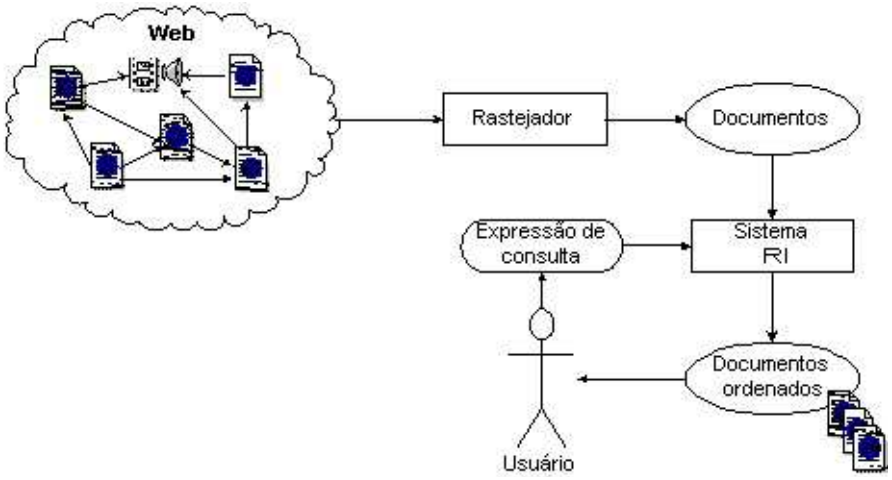


Figura 2-2 – Processo de recuperação de informações na *web*.

Arasu *et al.* (2001) afirma que a maioria dos algoritmos de recuperação de informação foram desenvolvidos para coleções pequenas de informações, tais como artigos de jornais ou catálogos de livros em uma livraria. A internet é massiva, mutante e está espalhada geograficamente em milhares de computadores. Vários fatores tornam o processo de recuperação de informações disponíveis na web mais complexo:

- Dados distribuídos: documentos espalhados em milhares de diferentes servidores *web*;
- Volatilidade dos dados: muitos documentos mudam ou desaparecem rapidamente, por exemplo, *links* quebrados;
- Dados redundantes e desestruturados: não há estrutura uniforme pré-definida para os documentos disponíveis e estima-se que cerca de 30% dos documentos sejam duplicados;

- Grandes volumes: bilhões de documentos separados;
- Qualidade dos dados: não há controle de edição, informações falsas ou erradas, textos mal escritos, entre outros;
- Dados heterogêneos: múltiplos tipos de mídia (imagens, vídeos, sons), linguagens e conjuntos de caracteres.

Nas próximas seções, analisaremos em detalhe a recuperação de informação na Web, ao descrever os mecanismos de busca e analisar o caso de dados geográficos.

2.2 Mecanismos de busca

Um mecanismo de busca na Web possui uma parte *on line* e outra *off line*. A parte *off line* é executada periodicamente pelo mecanismo e consiste em copiar sub-conjuntos da *web* para construir uma coleção de páginas que serão indexadas. A parte *on line* acontece sempre que uma consulta é executada. Consiste em selecionar documentos que podem satisfazer à consulta e organiza-los de acordo com uma estimativa de relevância (Castillo, 2004). A Figura 2-3 ilustra este processo.

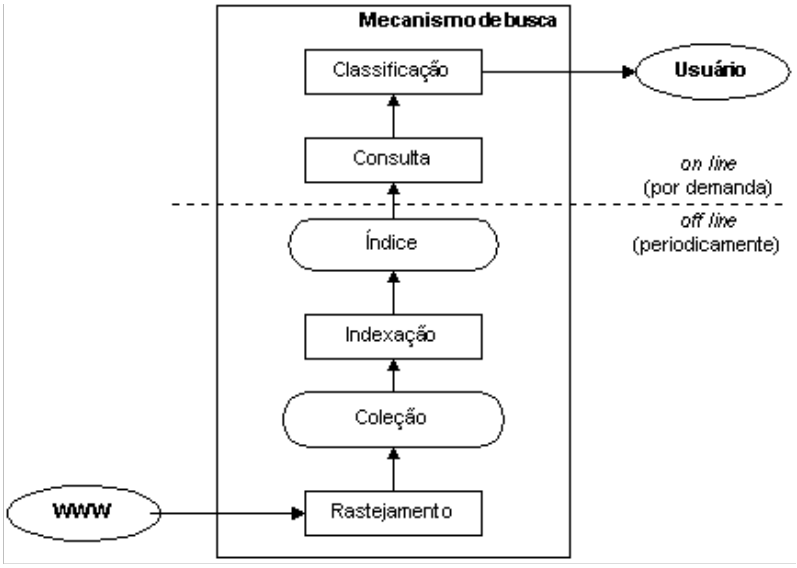


Figura 2-3 – Processos executados por um mecanismo de busca.

Adaptado de (Castillo, 2004)

A estrutura básica de um mecanismo de busca é composta por três componentes: rastejador, indexação e consulta e classificação.

2.2.1 Rastejador

O rastejador é um programa que percorre o ciberespaço de forma metódica e automatizada. Na literatura encontram-se diferentes denominações como *crawler*, indexador automático, *bot* e aranha (Kobayashi; Takeda, 2000). Utiliza a estrutura de *links* da internet para visitar os sítios e encontrar novas páginas interessantes. Os rastejadores são utilizados principalmente para copiar o conteúdo das páginas visitadas para serem posteriormente analisadas por um mecanismo de busca. Após análise, estas páginas são indexadas, o que permite pesquisas com velocidade e qualidade. O rastejador entende a internet como um grafo, onde os nós são recursos (páginas web e arquivos) localizados por URLs.

No início, o rastejador utiliza uma lista de URLs a ser visitadas. À medida que visita as URLs, ele identifica todos os *hyperlinks* na página e os adiciona à lista de URLs. As URLs que compõem a lista são visitadas de forma recursiva, obedecendo aos critérios de seleção, revisita, cortesia e paralelização.

A **política de seleção** define quais páginas devem ser copiadas. Devido à quantidade limitada de páginas que um rastejador consegue visitar (Lawrence; Giles, 1999), é importante que as páginas recuperadas sejam relevantes. Para tanto é fundamental estabelecer métricas de importância para a priorização de páginas. A importância de uma página é função de sua qualidade intrínseca, sua popularidade em termos de *links* ou visitas e suas URLs. Vários trabalhos propõem métricas de ordenamento (Cho *et al.*, 1998), (Najork; Wiener, 2001), (Boldi *et al.*, 2004), (Abiteboul *et al.*, 2003) e (Baeza-Yates *et al.*, 2005). Definir uma boa política de seleção é uma tarefa complexa. É preciso tratar informações parciais, pois o conjunto completo de páginas *web* é desconhecido durante o processo de rastejamento.

A **política de revisita** é necessária devido à natureza dinâmica da *web*. À medida que o rastejador percorre a *web*, vários eventos, tais como inclusões, atualizações e exclusões, modificam os recursos existentes. Para os mecanismos de busca, não detectar um evento

ocasiona cópias desatualizadas dos recursos. Cho (2000) definiu métricas para avaliar a idade (*age*) e o frescor (*freshness*) das páginas. A *idade* indica quão desatualizada está uma cópia local de uma página. O *frescor* é uma medida binária que indica se a cópia local é idêntica ou não à original. A Figura 2-4 apresenta as funções de *idade* e *frescor* de uma página *p* no tempo *t*.

A **idade** de uma página *p* no repositório em um tempo *t* é definido como:

$$A_p(t) = \begin{cases} 0 & \text{se } p \text{ não estiver modificado no tempo } t \\ t - \text{tempo de modificação de } p & \text{caso contrário} \end{cases}$$

O **frescor** de uma página *p* no repositório em um tempo *t* é definido como:

$$F_p(t) = \begin{cases} 1 & \text{se } p \text{ é igual à cópia local no tempo } t \\ 0 & \text{caso contrário} \end{cases}$$

Figura 2-4 – Funções de *idade* e *frescor* de uma página *p* no tempo *t*.

(Cho *et al.*, 2000)

Um dos objetivos do rastejador é manter a média de frescor tão alta quanto possível e a média da idade, baixa. Para manter uma média alta de frescor, o melhor método é ignorar as páginas que mudam freqüentemente. O rastejador desperdiça tempo ao tentar revisitar essas páginas em espaços curtos de tempo e, mesmo assim, ele não seria capaz de manter uma cópia atualizada delas (Cho; Garcia-Molina, 2003). Para manter baixa a média da idade, o melhor método é utilizar uma freqüência de acessos que aumente de acordo com a taxa de mudança de cada página. Cho (2003) propõe revisitar todas as páginas na coleção com a mesma freqüência para manter uma baixa idade e frescor das mesmas.

A **política de cortesia** indica como evitar sobrecarga nos sítios da *web*. A sobrecarga ocorre quando o rastejador executa múltiplas requisições por segundo ou quando copia arquivos grandes. Uma das alternativas para evitar o problema da sobrecarga é o protocolo de exclusão de robôs[3] (Koster, 1996). Embora o protocolo não estabeleça o intervalo entre as visitas, ele indica quais partes dos servidores *web* podem ser acessados. Intervalos de tempo entre os acessos variando de 1 a 60 segundos foram propostos por diversos pesquisadores

[3] Robôs são programas que percorrem páginas web recursivamente recuperando o conteúdo das páginas. Neste documento utilizamos o termo “rastejadores” para denominar o robô implementado.

(Koster, 1993), (Baeza-Yates; Castillo, 2002) e (Dill *et al.*, 2002). O rastejador MercatorWeb (Heydon; Najork, 1999) propôs uma política de cortesia adaptativa. Nela, o tempo entre os acessos para uma página específica é estabelecido como 10 *t*, sendo *t* o tempo em segundos que o rastejador requer para copiar um dado documento nessa página.

A **política de paralelização** estabelece como coordenar múltiplos rastejadores, para maximizar o número de recursos copiados e evitar que um recurso seja copiado mais de uma vez. Duas políticas de paralelização, dinâmica e estática, foram propostas por Cho (2002). Na dinâmica, um servidor central atribui novas URLs para diferentes rastejadores dinamicamente, cabendo ao servidor o balanceamento de carga de cada rastejador. Na estática, existe uma regra fixa pré-estabelecida que define como atribuir URLs novas aos rastejadores.

Alguns exemplos de rastejadores são: World Wide Web Worm (Mcbryan, 1994), Google Crawler (Brin; Page, 1998), WebCrawler (Pinkerton, 1994), CobWeb (Silva *et al.*, 1999), Mercator (Heydon *et al.*, 1999), WebFountain (Edwards *et al.*, 2001), Ubicrawler (Boldi *et al.*, 2004) e FAST Crawler (Risvik; Michelsen, 2002). O WebBase (Hirai *et al.*, 2000), WIRE (Baeza-Yates *et al.*, 2002) e WebSPHINX (Miller; Bharat, 1998) são exemplos de rastejadores de código aberto.

2.2.2 Indexação

O objetivo principal da indexação é construir um índice dos sítios visitados e capturados pelo rastejador. A indexação automática consiste em organizar as páginas *web* usando algoritmos de análise sintática. Para isto, utiliza algumas informações extraídas do corpo da página e de *metatags*[4] presentes no arquivo, incluindo:

- Conteúdo da página – informações mais precisas e texto completo;
- Descrição da página – informações sucintas que descrevem o conteúdo da página;

[4] *Metatags* são linhas de código HTML embutidas nas páginas *web*. Devem ser incluídas na área <**head**> de um documento HTML, iniciando logo após a *tag* <**html**>, e finalizando anteriormente à *tag* <**body**>.

- Texto dos *hyperlinks* – oferecem excelentes indícios semânticos do tópico para o qual apontam;
- Palavras-chave – informações que caracterizam o conteúdo de uma página;
- Título – informações sobre o título de uma página;
- Textos destacados – diferentes tipos, tamanhos e estilos de fontes demonstram textos que provavelmente possuem maior importância na descrição da página;
- Primeira frase – normalmente fornece informações essenciais para a descrição da página.

A indexação e formulação de consultas adotam diferentes técnicas para a normalização de textos: *tokenization*, *stopwords* e *stemming*. Para exemplificar essas técnicas, tomemos o texto apresentado na Figura 2-5 como exemplo.

Capítulo 1

Introdução

1.1. A política de revisita é necessária devido à natureza dinâmica da *web*. À medida que o rastejador percorre a *web*, vários eventos, tais como inclusões, atualizações e exclusões, modificam os recursos existentes.

Figura 2-5 – Exemplo de texto formatado.

A técnica de *tokenization* consiste em dividir o texto em uma seqüência de símbolos, em que cada símbolo é considerado uma palavra. Para tanto, remove a pontuação e caracteres especiais. Números podem ou não ser incluídos. Em alguns idiomas, tais como o português e o inglês, esta técnica é eficiente e relativamente simples, porém em outros idiomas, como o chinês, este processo torna-se mais complexo. O resultado desta técnica de normalização é uma versão pura de um texto completo. A Figura 2-6 apresenta o texto após a aplicação da técnica de *tokenization*.

capítulo 1 introdução 1 1 a política de revisita é necessária devido à natureza dinâmica da web à medida que o rastejador percorre a web vários eventos tais como inclusões atualizações e exclusões modificam os recursos existentes

Figura 2-6 – Texto após a técnica de *tokenization*.

Já *stopwords* são palavras que carregam pouca informação semântica, normalmente palavras funcionais ou siglas. No passado, os sistemas de recuperação informação não indexavam as *stopwords* para otimizar o espaço de armazenamento devido à sua alta freqüência. Atualmente, os mecanismos de busca utilizados indexam as *stopwords*. Vários estudos atestam que não existe um impacto real na eficiência do armazenamento das mesmas quando se trata de grandes bases de dados, como, por exemplo, a *web*. A Figura 2-7 apresenta o texto resultante após a remoção das *stopwords*.

capítulo 1 introdução 1 1 política revisita necessária devido natureza dinâmica web medida rastejador percorre web vários eventos inclusões atualizações exclusões modificam recursos existentes

Figura 2-7 – Texto após a remoção de *stopwords*.

A técnica de *stemming* é utilizada para extrair a origem morfológica das palavras. Converte palavras no plural para singular e formas verbais para o infinitivo, entre outras mudanças. Em mecanismos de busca genéricos, o *stemming* tem problemas devido à dependência do idioma e à ambigüidade das palavras que possibilitam a geração de mais de um radical. A Figura 2-8 apresenta o texto após a aplicação da técnica de *stemming*.

capitul 1 introdu 1 1 politic revisit necessari devid natur dinami web medi rastej percorr web vari event inclu atualiz excl modific recurso exist

Figura 2-8 – Texto após a aplicação da técnica de *stemming*.

Outras técnicas para normalizar textos incluem traduzir sinônimos, detectar expressões (por exemplo, ‘estado da arte’) e explicitar palavras com sentido ambíguo. Tais técnicas podem ser utilizadas, porém são mais complexas e acumulam uma taxa de erro significativa, que pode comprometer a precisão do processo de recuperação de informações.

Para indexar páginas, a maioria dos mecanismos de busca constrói um índice invertido composto por um vocabulário e uma lista de ocorrências. Todas as palavras que compõe uma página, incluindo *stopwords*, são organizadas em uma lista que mantém uma indicação de quais documentos possuem as palavras. Quando uma consulta é executada, o mecanismo utiliza as listas para as comparações com as palavras fornecidas na consulta.

O tamanho físico do índice depende da escolha de quais itens deverão compor o índice. Ao armazenar somente os identificadores dos documentos, gera-se um número reduzido de índices. Se a posição na qual as palavras aparecem em cada página também for armazenada, o número de índices aumenta consideravelmente. Neste caso, o índice será eficiente para responder consultas mais complexas, tais como frases exatas (Castillo, 2004).

A Figura 2-9 apresenta um exemplo de índice invertido com base em três documentos. Todas as palavras (incluindo artigos, preposições, pronomes etc.) foram consideradas e suas letras foram normalizadas para minúsculas. No exemplo, as consultas utilizaram o operador "e", porém outros operadores podem ser utilizados.

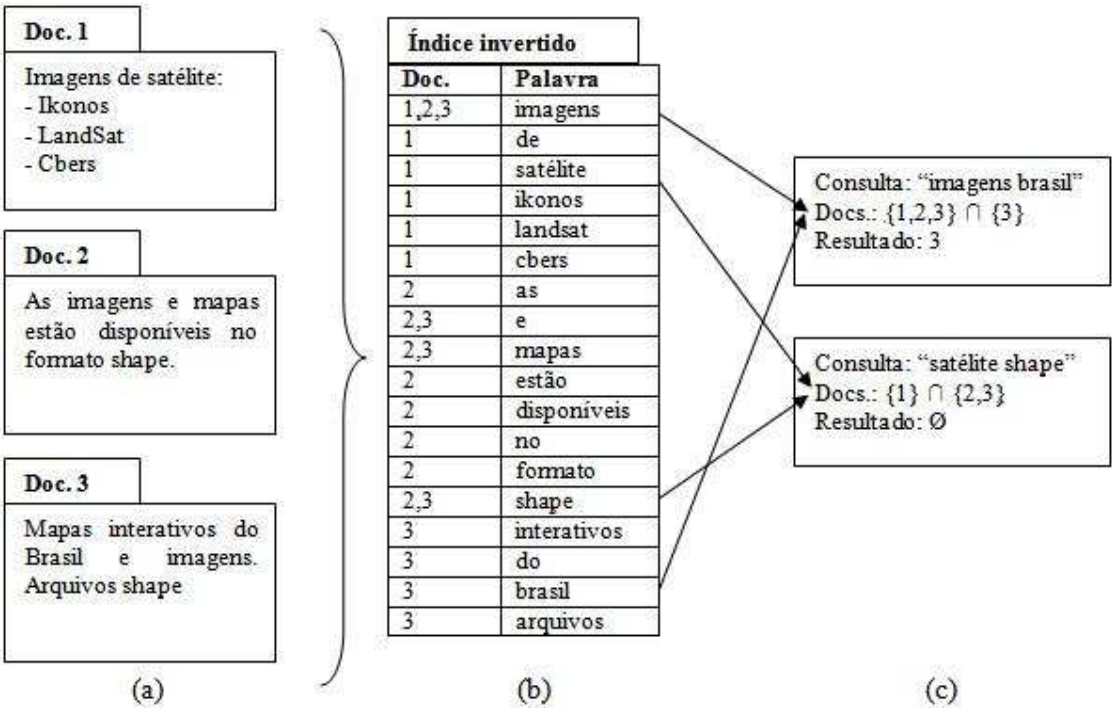


Figura 2-9 – Índice invertido de palavras: (a) documentos indexados; (b) índice construído (c) resultados apresentados às consultas.

2.2.3 Consulta e classificação

A principal finalidade de uma consulta é fornecer resultados relevantes com qualidade e velocidade. O processamento de uma consulta consiste em analisar as palavras ou expressões fornecidas pelo usuário e compará-las com o índice, a fim de encontrar respostas relevantes ao seu interesse. As expressões podem ser compostas por operadores “e”, “ou” e “não”.

Para estabelecer a ordem de demonstração dos resultados, o mecanismo utiliza um algoritmo de ordenação. Esse algoritmo tem diferentes critérios. Pode ser baseado na quantidade de vezes que a palavra aparece na página, pela classificação de *hubs* e autoridades (Kleinberg; Lawrence, 2001). Temos ainda algoritmos especiais baseados na estrutura de *links* da *web* como o PageRank (Page *et al.*, 1999) e o Hyperlink Vector Voting (HVV) (Li, 1998).

2.3 Medidas de qualidade em recuperação da informação

Para medir o desempenho de sistemas de recuperação de informações, seis diferentes critérios de avaliação foram julgados críticos pelos usuários: *recall*, *precision*, *effort*, *time*, *form of presentation* e *coverage* (Cleverdon, 1970). Dentre eles, as métricas de *abrangência (Recall)* e de *precisão (Precision)* são as mais utilizadas para avaliar a eficiência de sistemas de recuperação de informação.

A abordagem mais objetiva e compreensível considera três diferentes aspectos: os recursos usados na operação de recuperação, a quantidade de tempo e esforços gastos para obter a informação e a habilidade que o sistema possui para recuperar itens úteis (Raghavan *et al.*, 1989). É difícil obter todos os parâmetros relevantes para a medida. Desta forma, uma prática comum em investigações de pesquisa é concentrar principalmente em medidas que remetam à qualidade da produção da recuperação.

Para a definição de *abrangência* e *precisão*, Buckland (1994) baseou-se nas seguintes suposições: (a) classificação binária de relevância, na qual o item recuperável é classificado como “relevante” ou como “não relevante”; (b) a recuperação é vista como um processo expansivo, na qual o buscador aumenta sua *abrangência* continuamente.

Relevância é uma medida de quão bem um item atende à expectativa do usuário. A *relevância* é uma medida de difícil quantificação, devido às seguintes características (Lee, 2005):

- Subjetiva – depende do julgamento de um usuário específico;
- Situacional – relacionada às necessidades atuais do usuário;
- Cognitiva – depende da percepção e comportamento humano;
- Dinâmica – muda com o tempo.

A primeira suposição (a) adota uma matriz de recuperação, conforme demonstrado na Figura 2-10, que combina a classificação de itens recuperados e não recuperados, com a classificação de itens relevantes e não relevantes.

	Relevante	Não relevante	Total
Recuperado	$N_{rec \cap rel}$	$N_{rec \cap \overline{rel}}$	N_{rec}
Não recuperado	$N_{\overline{rec} \cap rel}$	$N_{\overline{rec} \cap \overline{rel}}$	$N_{\overline{rec}}$
Total	N_{rel}	$N_{\overline{rel}}$	N_{tot}

Figura 2-10 – Matriz de recuperação (Buckland *et al.*, 1994).

Para um dado conjunto de itens recuperados, a *abrangência* é definida como a proporção entre o número de itens relevantes recuperados e o número total de itens relevantes no sistema em questão.

$$Abrangência = \frac{N_{rec \cap rel}}{N_{rel}}$$

Para um dado conjunto de itens recuperados, *precisão* é definida como a proporção entre o número de itens relevantes recuperados e o número total de itens recuperados.

$$Precisão = \frac{N_{rec \cap rel}}{N_{rec}}$$

A *abrangência* é uma medida de efetividade na inclusão de itens relevantes no conjunto recuperado. A *precisão* é como uma medida de desempenho da recuperação. A *abrangência*

total pode ser alcançada examinando a base de dados completa, porém uma alta *abrangência* nem sempre é necessária, uma vez que os usuários comumente preferem que a busca retorne alguns itens relevantes.

O cenário ideal seria alcançar a totalidade de *abrangência* e de *precisão*. A Figura 2-11 apresenta a coleção completa de itens: os itens relevantes, os itens recuperados por um sistema (caracterizando a *abrangência*) e a intersecção dos relevantes com os recuperados (caracterizando a *precisão*).

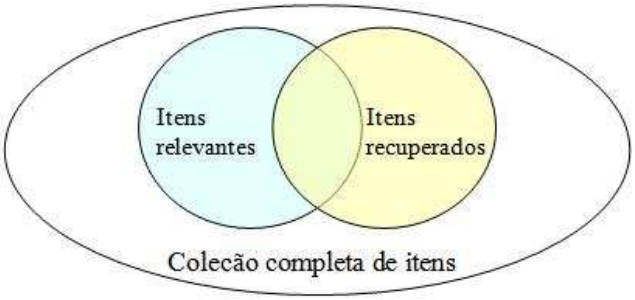


Figura 2-11 – Coleção completa de itens.

Abrangência e *precisão* são freqüentemente objetivos contraditórios na medida em que ao se desejar obter mais itens relevantes (aumentando o nível de *abrangência*), mais itens irrelevantes também são recuperados (diminuindo o nível de *precisão*). Estudos empíricos sobre o desempenho mostram uma tendência de declínio da *precisão* na medida em que a *abrangência* aumenta. Buckland (1994) relata que um sistema pode alcançar índices altos de *precisão* e *abrangência*, porém não simultaneamente. A Figura 2-12 demonstra o gráfico de compensação entre *abrangência* e *precisão*. O arco contínuo representa o comportamento real quando *abrangência* e *precisão* são aplicadas a uma coleção de itens e o arco tracejado demonstra o comportamento desejado ou ideal.

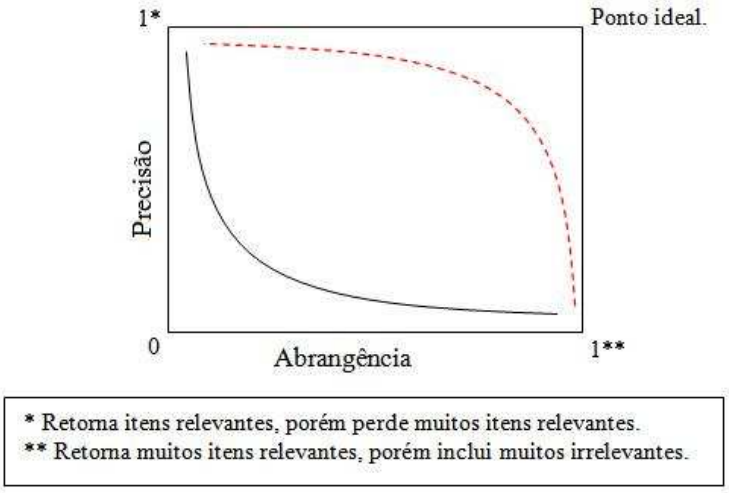


Figura 2-12 – Gráfico de compensação entre *abrangência* e *precisão*.

Desta forma, como podemos afirmar que o Sistema A é melhor que o Sistema B com base na determinação de *abrangência* e *precisão*? Pode-se afirmar que o Sistema A é melhor que o Sistema B se, em todos os pontos de *abrangência*, o valor da *precisão* de A for maior do que B. Caso isso não aconteça, as médias dos valores de *precisão* para valores de *abrangência* selecionados são calculadas e comparadas. *Abrangência* e *precisão* são medidas depois que o sistema ordena os itens nas coleções que são apresentadas a partir de uma consulta do usuário. Esta ordem representa a classificação que o sistema faz sobre o quanto cada item recuperado atende às expectativas do usuário.

Em grandes bases de dados ou sistemas com capacidade de recuperação limitada pode haver vantagens na recuperação dos itens em duas etapas: recuperação inicial objetivando um alto índice de abrangência, seguido por buscas mais detalhadas no conjunto de itens recuperado inicialmente. Esta estratégia pode melhorar a *precisão* e *abrangência*, porém a inversão entre eles permanece (Buckland *et al.*, 1994).

2.4 Arquivos geográficos na internet

A disponibilidade de arquivos geográficos na *web* aumenta concomitantemente com a expansão da internet, a efetiva utilização de dados geográficos por diferentes profissionais na execução de suas atividades e o conseqüente aumento na produção desses dados.

Atualmente, reconhecidos produtores de dados[5] compartilham os dados geográficos na *web* e vários fornecedores de programas para SIG oferecem alternativas para acesso de dados geográficos através da *web*. Casanova et al (2005) apresentam diferentes abordagens para a disseminação de dados geográficos na *web*:

(a) A disseminação direta;

(b) As bibliotecas digitais de informações geográficas;

(c) Infra-estruturas para dados espaciais;

A disseminação direta usa características gráficas típicas dos navegadores, complementadas por recursos adicionais. Inclui: (a) a apresentação de mapas estáticos em formato de imagem inseridos diretamente nos documentos HTML; (b) mapas gerados a partir de formulários, onde é possível informar dados sobre uma determinada região geográfica e o sistema apresenta um mapa final em formato de imagem; (c) mapas gerados a partir de um mapa-chave, onde o usuário pode definir a área a ser visualizada com mais detalhes; e (d) a transmissão de objetos geográficos com representação vetorial, permitindo maior interatividade com o usuário.

As bibliotecas digitais de informações geográficas ou centro de dados geográficos permitem a coleta, armazenamento, processamento, acesso e distribuição de dados ou programas através de uma rede de uma rede privada ou uma rede pública (Câmara *et al.*, 1996a). Dentre os exemplos citados por Casanova et al (2005) estão a Alexandria Digital Library (Frew *et al.*, 2000), a Maine Library of Geographic Information e o GeoConnections Discovery Portal.

Infra-estruturas de dados espaciais permitem acesso à informação geográfica a partir de catálogos de acervos de informação. Para acessar os dados, o usuário não precisa conhecer detalhes de armazenamento nem a estrutura física dos dados (que está encapsulada).

Além das formas discutidas acima, existe uma grande quantidade de arquivos geográficos disponíveis na *web* que são postados diretamente nos sítios dos produtores de dados sem

[5] No contexto deste trabalho, produtores de dados são profissionais ou empresas que produzem dados geográficos, por exemplo, o INPE, U.S. Census Bureau, IBGE.

nenhum tratamento especial. Estes arquivos estão disponíveis para cópia e, muitas vezes, encontram-se compactados para facilitar sua distribuição. Dentre os formatos mais difundidos, estão os formatos *GeoTIFF* (Ritter; Ruth, 1997) e *Shapefile* (ESRI, 1998).

O *GeoTIFF* é uma extensão do formato TIFF – *Tagged Image File Format* específica para dados matriciais geográficos. É recomendada pela OGC como o formato padrão para o intercâmbio de dados matriciais (OGC, 1999). O formato TIFF foi desenvolvido com a finalidade de se tornar um padrão para imagens no âmbito comercial. É um formato binário que utiliza *tags* (marcadores) para armazenar informações básicas sobre a imagem que representa. Os marcadores básicos contêm informações sobre o número de linhas e de colunas, o número de componentes e o número de *bits* por *pixel* da imagem. O formato *GeoTIFF*, por sua vez, possui um conjunto específico de marcadores para descrever as informações cartográficas e geodésicas associadas à imagem TIFF. A Figura 2-13 apresenta um arquivo TIFF georreferenciado com seus parâmetros para o mapeamento das coordenadas de imagem com as coordenadas terrestres.

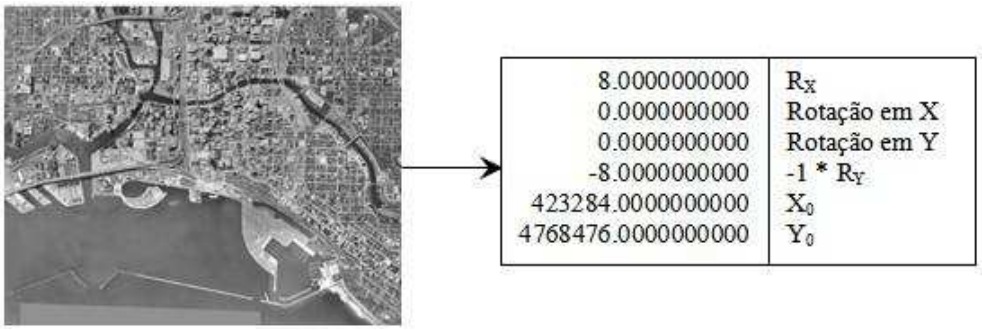
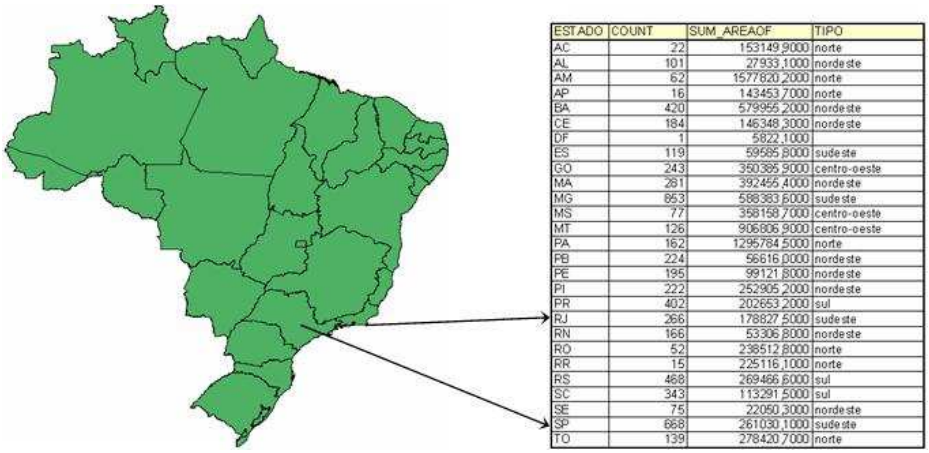


Figura 2-13 – Arquivo TIFF e os parâmetros de mapeamento.

O formato *ShapeFile* armazena geometria não-topológica e informações de atributos de *features*[6] espaciais em um conjunto de dados. A geometria é armazenada como uma figura juntamente com um conjunto de coordenadas vetoriais. O *shapefile* suporta pontos, linhas, e polígonos e os atributos são armazenados em arquivos no formato dBASE. Para cada atributo armazenado há um relacionamento um-para-um com um *shape* (ESRI, 1998).

[6] ***Feature:*** *representação* de uma característica geográfica que possui tanto uma representação especial relacionada à uma forma quanto um conjunto de atributos.

Um *ShapeFile* consiste em um arquivo principal, um arquivo de índice e uma tabela no formato dBASE. O arquivo principal possui registros de tamanho variável, e cada registro descreve um *shape* com uma lista de seus vértices. O arquivo de índice contém a organização do arquivo principal, permitindo o acesso direto aos dados. A tabela dBASE contém os atributos das *representações* com um registro por *representação*. A Figura 2-14 apresenta o arquivo principal e a respectiva tabela com os registros relacionados.



ESTADO	COUNT	SUM_AREAOF	TIPO
AC	22	153149,9000	norte
AL	101	27933,1000	nordeste
AM	62	1577820,2000	norte
AP	16	143453,7000	norte
BA	420	579955,2000	nordeste
CE	184	146348,3000	nordeste
DF	1	5822,1000	
ES	119	59585,8000	sudeste
GO	243	350385,9000	centro-oeste
MA	281	392456,4000	nordeste
MG	853	588383,6000	sudeste
MS	77	358158,7000	centro-oeste
MT	126	906806,9000	centro-oeste
PA	162	1295784,5000	norte
PB	224	56616,0000	nordeste
PE	195	99121,8000	nordeste
PI	222	252905,2000	nordeste
PR	402	202653,2000	sul
RJ	266	178827,5000	sudeste
RN	166	53306,8000	nordeste
RO	52	238512,8000	norte
RR	15	225116,1000	norte
RS	468	269466,6000	sul
SC	343	113291,5000	sul
SE	75	22050,3000	nordeste
SP	668	261030,1000	sudeste
TO	139	278420,7000	norte

Figura 2-14 – Arquivo *shape* com o respectivo arquivo dBase.

O arquivo principal (*.shp) possui um cabeçalho de arquivo com tamanho fixo, composto por registros com tamanhos variáveis. O arquivo de índice (*.shx) possui um cabeçalho idêntico ao do arquivo principal ao qual está relacionado. Este cabeçalho é composto por registros de tamanho fixo, campos que armazenam o *OffSet* (posição do registro em questão no arquivo principal, a partir do inicio do arquivo) e o tamanho do registro no arquivo principal.

O arquivo dBASE (*.dbf) contém um conjunto de atributos das *representações* ou atributos-chave que estabelecem relacionamentos com outras tabelas. O formato segue o padrão para arquivos DBF usado por aplicações Windows e DOS. As tabelas podem armazenar quaisquer conjuntos de campos, desde que atendam a três requisitos:

- O nome do arquivo deve obedecer à convenção de nomes e deve ter o mesmo prefixo que os arquivos principal e índice;

- A tabela deve conter um registro por *representação*;

- A ordem dos registros deve ser a mesma das *representações* no arquivo principal.

2.5 Mecanismos de busca no contexto geográfico

As questões que envolvem a recuperação de informações geográficas na *web* com a utilização de mecanismos de busca especializados estão sendo tratadas por diversas pesquisas e projetos, alguns dos quais são apresentados nesta seção.

O projeto *Spirit* consiste em um mecanismo de busca que inclui uma interface do usuário, funcionalidades para a manutenção e recuperação de informações, uma ontologia geográfica e procedimentos para classificação de relevância e extração de metadados de documentos. O Spirit é uma extensão do mecanismo de busca textual GLASS. A principal modificação foi a introdução da indexação espacial dos índices dos documentos *web* e mecanismos para busca por contexto geográfico presente nos documentos *web* (Jones *et al.*, 2004). Adicionalmente às técnicas utilizadas para a extração de contexto geográfico dos documentos *web*, o Spirit detecta características em conjuntos de dados geográficos para incrementar a ontologia geográfica utilizada.

O Geo-Tumba (Silva *et al.*, 2004) usa o escopo geográfico das páginas *web* para encontrar páginas relevantes com conteúdo geográfico. Para agrupar resultados relacionados a um mesmo escopo ou para classificar os resultados de acordo com a proximidade com uma dada localização, ele considera a similaridade espacial entre as páginas. O projeto apóia-se em três componentes principais: a formulação de consultas, elucidação do contexto espacial das consultas e apresentação dos resultados.

O projeto NetGeo (Moore *et al.*, 2000) utiliza locais geográficos no contexto de páginas *web* para colecionar informações de múltiplas origens e atribuir uma latitude/longitude mais provável a partir da localização do endereço IP do servidor. Ele considera que há um relacionamento entre a localização dos servidores de páginas web e o conteúdo destas páginas.

O Windows Live Local (http://local.live.com/) permite que usuários encontrem locais nos EUA ou Canadá por meio de um determinado endereço. As consultas são possíveis por

nome ou categoria. Os resultados da busca são demonstrados em um mapa no qual o usuário pode escolher a escala da visualização.

O Google Local (http://local.google.com) é um mecanismo de busca especializado em busca por regiões geográficas. Sua principal funcionalidade é permitir que usuários encontrem informações específicas de locais específicos nos EUA. Para alimentar sua base de dados alfanumérica, compila informações de diferentes origens (tais como os resultados de seus próprios rastejadores, dados submetidos por proprietários de comércios e outras fontes públicas disponíveis como, por exemplo, diretórios de páginas amarelas) e sua base gráfica é composta por mapas produzidos pelas agências NAVTEQ e TeleAtlas e imagens de satélite das empresas DigitalGlobe e EarthSat.

O Google Earth (http://earth.google.com) também combina técnicas de mecanismos de busca, imagens de satélite, fotos aéreas e mapas. Possui ferramentas para a manipulação e o reconhecimento de uma localização retornada em uma consulta, tais como visualização tridimensional, medidas de distâncias e áreas, desenho de linhas e formas e visualização de comércios próximos de uma localização selecionada.

O mecanismo Mirago (www.mirago.co.uk) executa busca em regiões da Inglaterra. A partir de uma consulta executada por um usuário, apresenta os resultados destacando áreas específicas no mapa ou selecionando uma região de uma lista predefinida. Assim como no Google Local, a funcionalidade geográfica, no mecanismo Mirago, prioriza a limitação de buscas por regiões.

Laender (2004) utiliza dados geoespaciais extraídos de páginas *web* em um SIG urbano que utiliza endereços como base para a integração de dados de diferentes fontes na *web* com imagens de alta resolução. Para tanto, desenvolve uma técnica que utiliza rastejadores para coletar dados de interesse na internet, extrai indicadores geográficos com análise sintática da página, converte os dados para o formato XML, geocodifica os endereços para obter as coordenadas para a atualização do banco de dados GIS e, finalmente, integra as informações dentro de várias fontes de dados GIS.

CAPÍTULO 3

ANATOMIA DE UM MECANISMO DE BUSCA ESPECIALIZADO EM ARQUIVOS GEOGRÁFICOS

O capítulo anterior apresentou os princípios gerais dos mecanismos de busca na Internet. Indicamos que, para áreas específicas, é possível desenvolver um buscador especializado. Neste e nos próximos capítulos, descrevemos um mecanismo de busca especializado em dados geográficos, chamado *GeoDiscover*. Neste capítulo descreveremos a arquitetura implementada no *GeoDiscover*. Apresentaremos duas visões da anatomia do sistema: a visão conceitual, que trata da organização dos processos de descoberta, classificação, indexação e recuperação de dados, e a visão estrutural, que trata da estrutura física dos servidores, colaboradores e aplicações desenvolvidas para suportar todo o mecanismo. A visão conceitual será detalhada nos capítulos seguintes. A visão estrutural é discutida neste capítulo, com especial atenção ao processo de gerenciamento das aplicações e o desempenho do *GeoDiscover*.

3.1 Arquitetura proposta

As principais características do GeoDiscover são:

- O emprego de arquitetura distribuída para a descoberta, captura e análise de páginas *web*;
- Algoritmos integrados para a análise do contexto do sítio (texto-âncora e texto-âncora estendido) em busca de indícios de arquivos geográficos e análise do conteúdo dos arquivos geográficos;
- Metodologia para a classificação dos resultados retornados às consultas; e

O buscador tem três módulos principais: o módulo de descoberta e captura de arquivos geográficos que é desempenhado por usuários colaboradores; o módulo de classificação e indexação que é desempenhado em servidores próprios; e o módulo de consulta e

visualização que possui interface amigável e é acessado por navegadores *web*. Os módulos desenvolvidos estão apresentados na Figura 3-1 e serão detalhados nos capítulos seguintes.

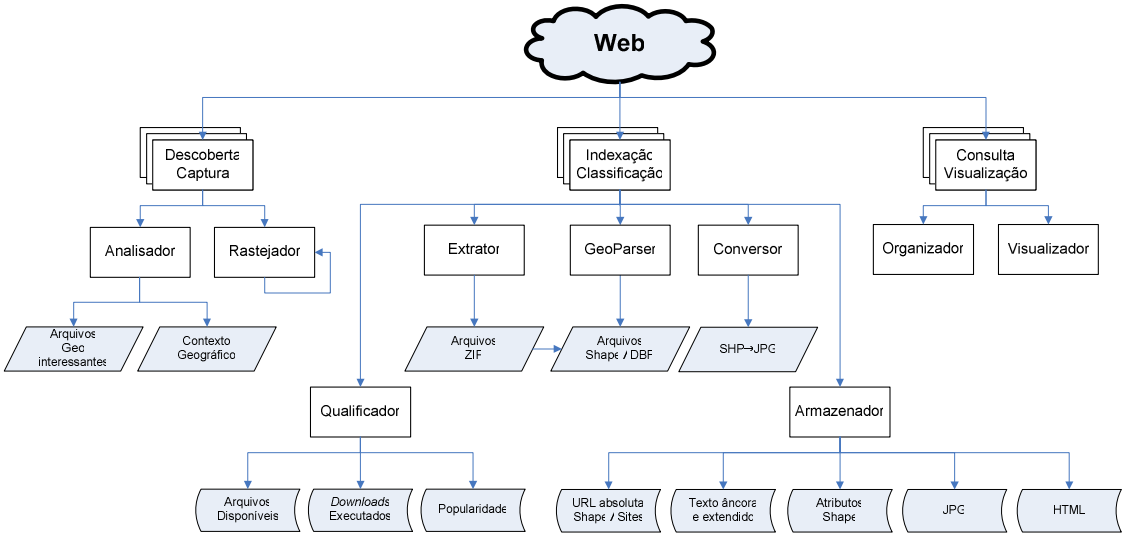


Figura 3-1 – Visão conceitual do mecanismo de busca

Para garantir eficiência nos processos de descoberta, classificação, indexação e recuperação dos arquivos geográficos, o *GeoDiscover* permite processamento distribuído com um servidor de aplicações centralizado com *web services* responsável pelo gerenciamento das requisições dos clientes. Os módulos foram divididos fisicamente em duas arquiteturas principais: usuários colaboradores e servidores. Usuários colaboradores são clientes que auxiliam no rastreamento de páginas *web* e no reconhecimento de páginas que possam conter arquivos geográficos[7]. O protótipo atual utiliza três servidores (Servidor de *Web Services* (WS), Servidor de Banco de Dados (BD) e Servidor de Aplicações (SA)), os quais são responsáveis por:

- Possibilitar a comunicação entre colaboradores, servidores e usuários;
- Gerenciar a distribuição de URLs que serão visitadas;
- Receber, organizar e armazenar o conteúdo HTML das páginas geo-interessantes;
- Receber, organizar e armazenar as URLs e metadados de arquivos geográficos;

[7] Neste trabalho, as páginas que possam conter arquivos geográficos serão denominadas “páginas geo-interessantes.”

- Executar processos de extração e conversão de arquivos geográficos;
- Executar a análise sintática sobre os arquivos geográficos;
- Disponibilizar uma interface para que o usuário possa executar consultas na *web*;
- Disponibilizar uma interface para a visualização dos resultados obtidos.

A Figura 3-2 explicita a interação existente entre os módulos definidos na visão conceitual e a arquitetura efetivamente adotada na estruturação do *GeoDiscover*.

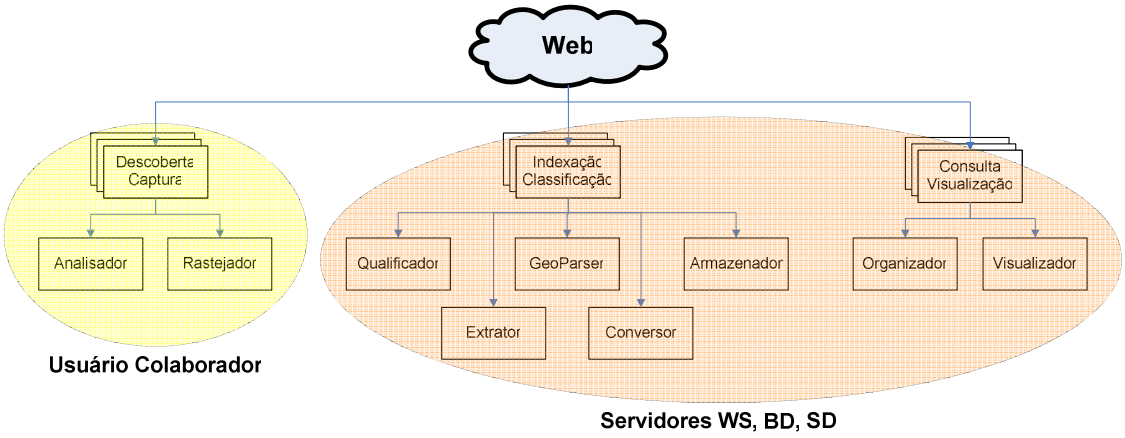


Figura 3-2 – Interação da visão conceitual e da estrutura física do *GeoDiscover*

3.2 Visão estrutural dos Servidores do *GeoDiscover*

O servidor de *Web Services* (WS) é responsável pela comunicação e pelo transporte de dados entre o módulo gerenciador dos colaboradores e os servidores de Banco de Dados (BD) e de Aplicações (SA), e entre os usuários e o BD durante a execução de consultas e recuperação de dados. O tráfego de dados entre os servidores, colaboradores e usuários acontece utilizando a linguagem de marcação estendida XML. Dessa forma, computadores protegidos por *firewall* e *proxy* também podem utilizar o programa.

O servidor de Banco de Dados armazena os dados capturados de páginas geo-interessantes. Dentre os dados armazenados estão: título da página, descrição da página, palavras-chave, texto completo da página, texto-âncora, URLs, data da última visita do rastejador, nome,

tamanho e tipo dos arquivos encontrados. Esses dados são utilizados para a classificação das páginas geo-interessantes, para as consultas executadas pelos usuários e para alimentar a lista de URLs que serão visitadas pelos colaboradores.

Os arquivos geo-interessantes não são armazenados permanentemente, pois devido ao tamanho físico dos mesmos e à quantidade de arquivos coletados na *web*, seriam necessários investimentos significativos em dispositivos de armazenamento. Arquivos geo-interessantes ocupam uma quantidade significativa de espaço, na ordem de *megabytes*. Considerando a grande quantidade de arquivos indexados pelo GeoDiscover, o espaço para armazenamento pode exceder rapidamente à escala dos *terabytes*, tornando inviável a utilização de um servidor para essa tarefa. Dessa forma, estes arquivos são armazenados temporariamente para a análise sintática e extração de metadados. Após passarem por tais processos, os arquivos são excluídos.

O Servidor de Aplicações (SA) é conectado fisicamente ao servidor BD por uma rede interna. Ele monitora os caminhos de arquivos geográficos incluídos no servidor de BD e inicia o processo de *download* para cada inclusão. Ele busca o arquivo em seu local original, captura-o e armazena-o em um diretório. Para tratar arquivos em formatos compactados, foi implementada uma rotina para descompactar e armazenar os arquivos descompactados em diretórios específicos. A análise sintática dos arquivos geográficos é executada no SA. Consiste em extrair metadados do arquivo *dbf* associado ao arquivo *shape*. A metodologia proposta para a análise sintática sobre os arquivos *bdf* será discutida detalhadamente na seção 3.8.4. Os metadados extraídos são enviados para o BD para armazenamento e são utilizados durante o processo de consulta do usuário.

Para facilitar a identificação de quais arquivos atendem aos critérios de consulta estabelecidos pelo usuário, para cada item da lista de arquivos *shape* retornados, são fornecidas informações adicionais do produtor de dados (título, descrição e URL da página) e informações do arquivo *shape* (nome do arquivo, tipo, data da extração, tamanho e quantidade de *downloads* executados a partir do *GeoDiscover*). As Figuras 3-3 e 3-4 apresentam os metadados de um produtor de dados e de um arquivo *shape*, respectivamente:

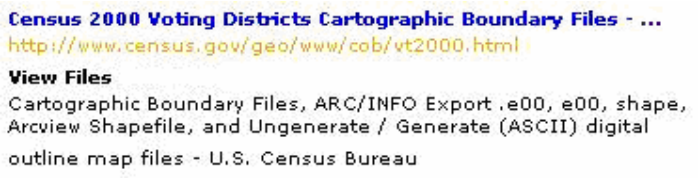


Figura 3-3 – Metadados do produtor de dados.

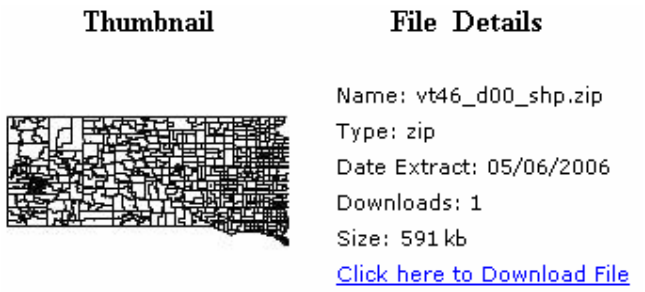


Figura 3-4 – Detalhes dos metadados de um arquivo shape.

Outra função desempenhada pelo SA é a geração de miniaturas, também denominadas *thumbnails*, a partir dos arquivos *shape* encontrados e indexados, conforme exemplificado na Figura 3-5. Os *thumbnails* são armazenados no servidor de BD e utilizados na composição da interface de resultados de consulta do usuário, permitindo a pré-visualização do arquivo que está sendo apresentado.

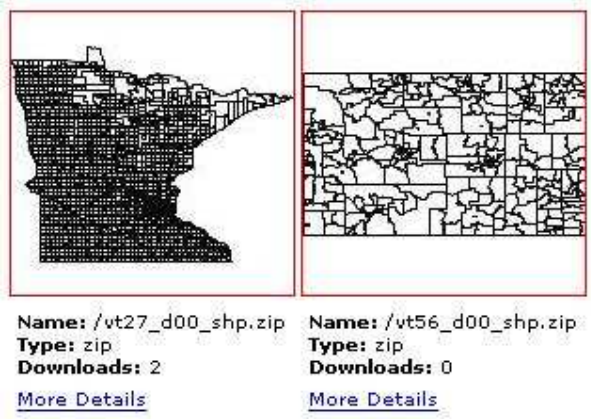


Figura 3-5 – *Thumbnails* gerados a partir de arquivos *shape* indexados.

Finalmente, cabe ao SA executar as rotinas para a qualificação dos produtores de dados. Essa qualificação é utilizada para ordenar os resultados de consultas do usuário.

3.3 Usuários colaboradores

O *GeoDiscover* utiliza em sua arquitetura usuários colaboradores[8] que auxiliam no processo de descoberta de arquivos geográficos. A decisão de estruturar o *GeoDiscover* para o processamento distribuído, utilizando usuários colaboradores para auxiliar no processo de rastreamento e análise sintática de páginas *web*, foi motivada pela crescente aceitação dos internautas em projetos virtuais colaborativos, pela capacidade de crescimento sustentável do mecanismo de busca à medida que novos usuários colaboradores inserem-se ao projeto e pela redução de investimentos necessários para a aquisição e manutenção de infra-estrutura.

Muitos projetos que utilizam o tempo ocioso de computadores de usuários conseguem agregar milhares de colaboradores, aumentando significativamente sua capacidade. Dentre esses projetos destacam-se o BOIC - Berkeley Open Infrastructure for Network Computing ((Anderson, 2004)), Seti@Home ((Anderson *et al.*, 2002), Predictor@Home ((Taufer *et al.*, 2005)) e Globus Project ((Foster; Kesselman, 1998)).

No *GeoDiscover*, os colaboradores contribuem no rastreamento do ciberespaço, procurando por arquivos geográficos e coletando páginas *web*, e na execução da análise sintática das páginas coletadas em busca de evidências de arquivos geográficos. A Figura 3-6 apresenta o fluxo de atividades executadas pelos colaboradores.

O processo para a descoberta de dados geográficos é iniciado quando o computador de um colaborador está ocioso. O módulo gerenciador do colaborador, denominado geo-colaborador, solicita uma lista de URLs para serem visitadas. O servidor de *Web service* (WS) envia um lista ordenada de URLs para o colaborador, que dispara o rastejador para encontrar os sítios indicados e recuperar o conteúdo HTML das páginas encontradas. O servidor WS também envia um vetor com o código *hash* das palavras geo-interessantes. Para a análise do conteúdo HTML das páginas *web* foi criado um repositório de palavras geo-interessantes que serão identificadas no contexto da página.

[8] Nesta tese utilizamos o termo "colaborador" para designar os computadores de usuários que auxiliam no processo de descoberta de páginas geo-interessantes.

Ao encontrar a página *web*, o rastejador verifica a existência de protocolos de exclusão de robôs[9]. Não havendo restrições para a indexação da página, o processo de *download* é iniciado e o conteúdo HTML é transferido para o módulo geo-colaborador. O geo-colaborador recebe os arquivos HTML e inicia a análise do contexto em busca de conteúdo geográfico.

Finalizada a análise sintática, e caso encontre páginas com indícios de arquivos geográficos ou encontre explicitamente arquivos geográficos, o geo-colaborador envia todo o conteúdo da página para o servidor WS. Caso contrário, enviará para o servidor WS somente as URLs encontradas na página para que possam ser visitadas posteriormente. No final do processo, o conteúdo HTML capturado é descartado e o ciclo é reiniciado com o rastreamento de uma nova página.

A análise sintática dos conteúdos HTML adotada pelo *GeoDiscover* é um processo complexo que inclui a identificação de contexto geográfico na página, o respeito ao código de ética dos rastejadores, a conversão de URLs relativas em URLs absolutas e o envio de dados para o servidor WS. A análise sintática será discutida em detalhes na seção 3.6.

Para se tornar um colaborador, o usuário necessita instalar em seu computador um módulo que gerencia as tarefas descritas acima. O módulo geo-colaborador conta com uma interface gráfica, que permite ao usuário acompanhar os processamentos desenvolvidos em seu computador quando a aplicação está sendo executada. Está disponível para instalação no sítio do *GeoDiscover* (www.geodiscover.org).

[9] O arquivo Robots.txt explicita regras de indexação do sítio que ele está presente e também regras para a utilização dos *links* para os quais o sítio aponta (www.robotstxt.org)

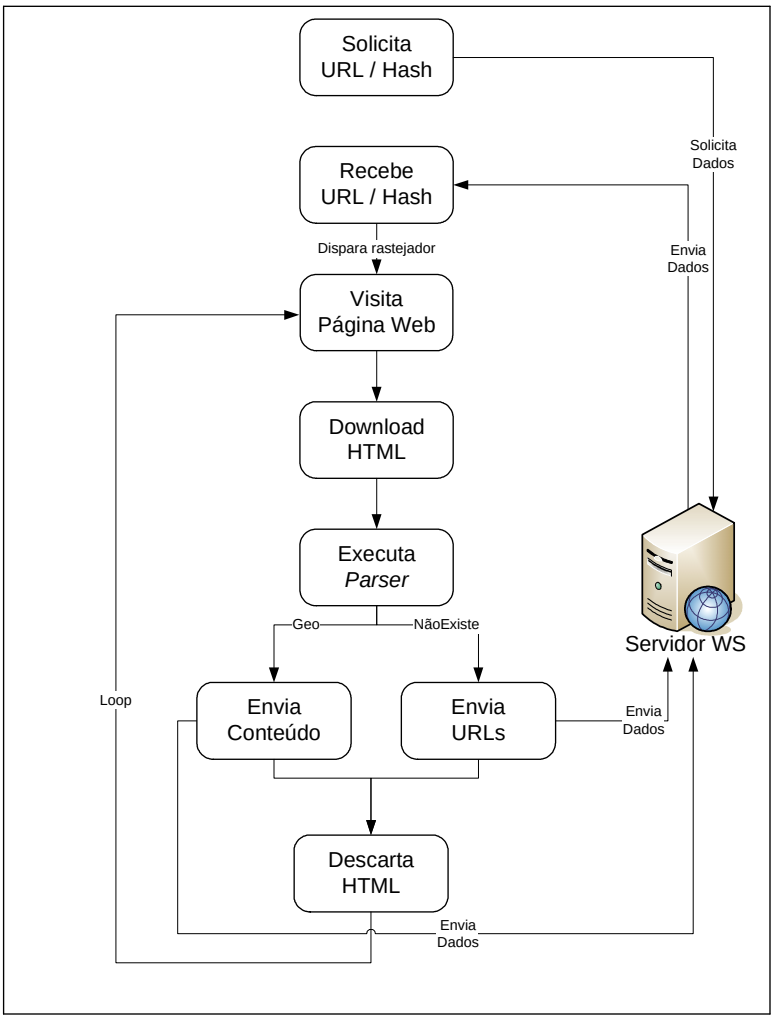


Figura 3-6 – Fluxo de tarefas do usuário colaborador.

3.4 Descoberta de arquivos geográficos

Como encontrar arquivos geográficos disponíveis na web? Similarmente como acontece em artigos científicos (onde o formato PDF é o predominante), produtores de dados geográficos usualmente distribuem seus dados em formatos que possibilitem sua leitura por diferentes sistemas utilizados pelos usuários. Podemos citar como exemplo os formatos *spr* do *software Spring* (Câmara *et al.*, 1996b) e o formato *shape (ESRI, 1998)* do *software ArcView*. O protótipo atual do *GeoDiscover* encontra e indexa arquivos *shape*.

Porém um dos grandes desafios para encontrar arquivos geográficos na *web* está na forma com que os produtores de dados disponibilizam seus arquivos. Muitos produtores

compactam os arquivos a fim de reduzir seu tamanho para otimizar o espaço de armazenamento e facilitar o processo de *download*. Frente a essa realidade, em muitos casos, os arquivos *shape* não estão disponíveis no formato *shp*, mas sim em formatos compactados, como por exemplo, *zip*. Em uma pesquisa executada em 47 sites de produtores de dados selecionados aleatoriamente na *web*, foram recuperados 2.677 arquivos *shape*. Destes, 16,9% estavam na extensão nativa do formato shape (*.shp, *.shx e *.dbf) e 62% estavam compactados no formato *.zip. A Figura 3-7 apresenta um gráfico com os formatos recuperados; os formatos que apresentaram percentuais inferiores a 1 foram agrupados e rotulados como *outros*.

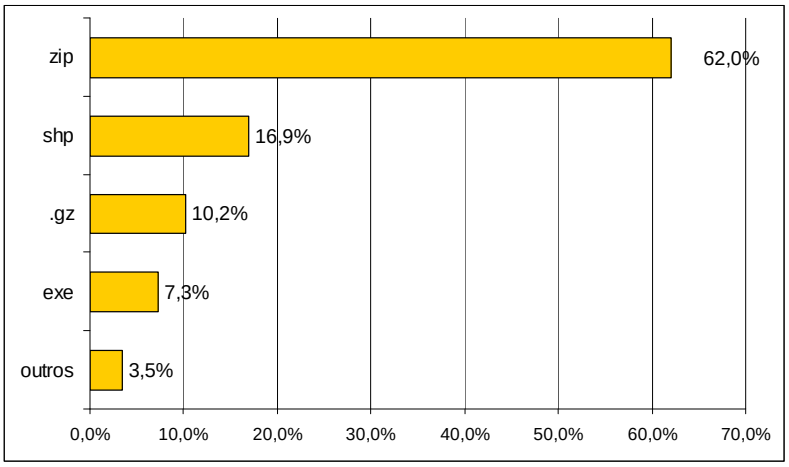


Figura 3-7 – Formatos de disseminação de arquivos *shapefile*.

Dessa forma, não basta especializar o rastejador para encontrar arquivos com extensão *shp*, pois esta solução reduziria drasticamente a quantidade de arquivos retornados. Por outro lado, capturar e analisar todos os arquivos compactados disponíveis na *web* traria sérios problemas de desempenho ao sistema e limitaria a cobertura da *web*. Para analisar os arquivos é necessário capturá-los e processá-los individualmente, despendendo uma quantidade de tempo significativa. Como alternativa, sabe-se que uma característica interessante dos arquivos *shape* disponíveis em formatos compactados é que parte significativa deles possui a palavra *shape* ou as letras *shp* na composição do nome do arquivo ou na URL que aponta para ele. Como exemplos, há arquivos com nomes de *PC_FLU_Shape_020106.zip*, *VT01D00SHP.zip* e *http://ross.urs-tally.com/docs/shapefiles/tracklines.zip*. Uma pesquisa executada a partir de 100 arquivos

shape compactados disponíveis na *web*, escolhidos aleatoriamente, mostrou que 42% deles apresentavam os conjuntos de caracteres *shape* ou *shp* na composição de seu nome ou de sua URL.

Para vencer este desafio e ampliar a quantidade de arquivos recuperados, o *GeoDiscover* utiliza uma metodologia que combina diferentes técnicas para descobrir dados geográficos:

- Busca explícita de arquivos pelo formato *shp*;
- Busca explícita de arquivos compactados que contenham na composição de seu nome os conjuntos de caracteres *shape* ou *shp*;
- Busca explícita de arquivos compactados que contenham na composição de sua URL os conjuntos de caracteres *shape* ou *shp*;
- Identificação de conteúdo com indícios geográficos em páginas *web* que apontem para arquivos compactados. Com isto recuperamos apenas arquivos compactados cuja probabilidade de serem geográficos seja alta.

As buscas explícitas de arquivos pelo formato *shp* e por arquivos compactados são executadas totalmente pelo rastejador, que é direcionado para verificar o formato desejado. Nos casos onde há a necessidade de analisar o nome dos arquivos e suas URLs, é necessário que o analisador sintático interaja no processo. Para ampliar a busca para outros formatos específicos (spr, geotiff etc), basta ajustar o rastejador. A identificação de conteúdo com indícios geográficos em páginas *web* que apontem para arquivos compactados é mais complexa e necessita da interação do rastejador e do analisador sintático de conteúdos *HTML*.

3.5 Rastejador

O rastejador é uma aplicação que vasculha o ciberespaço de forma metódica e automática utilizando a estrutura de *links* da *web (Kleinberg et al., 2001)*. Ele percorre o ciberespaço para descobrir novos recursos e indexar os documentos. Para recuperar os documentos dos servidores, o rastejador utiliza o protocolo de transferência de hipertextos (HTTP) da *web* (Cheong, 1996). O processo de rastejamento é uma atividade complexa, pois demanda a

interação com milhares de servidores *web* (Brin *et al.*, 1998). Dentre as principais dificuldades encontradas no processo de descoberta de páginas e arquivos disponíveis na *web*, destacam-se o grande volume de dados disponível e a grande quantidade de mudanças, caracterizadas por novas inserções, atualizações e exclusões.

Os rastejadores podem ser direcionados a nichos específicos dentro da coleção de recursos disponíveis na *web.* Podem ser direcionados para recolher endereços de correio eletrônico e telefones, ou para recolher arquivos específicos por meio da especialização em formatos pré-estabelecidos como o de imagens, som e documentos. Os rastejadores do *GeoDiscover* são especializados em arquivos com extensões *shp* e *zip*.

Para aumentar a abrangência da indexação de páginas *web*, o protótipo desenvolvido adota uma arquitetura distribuída de rastejadores que são executados nos computadores dos usuários colaboradores. Desta forma, as tarefas desempenhadas pelo rastejador são controladas pelo módulo geo-colaborador em sintonia com o Servidor de Aplicações. Para disparar um rastejador é necessário informar uma URL para ser visitada. O módulo geo-colaborador recebe uma lista de URLs que deverão ser visitadas. O rastejador utiliza cada URL da lista para visitar a página e enviar seu conteúdo HTML juntamente com dados da visita (data, horário e status) para o geo-colaborador. Antes de copiar o conteúdo da página, o rastejador verifica se o tempo resultante da última revisita com a data atual é maior que o tempo mínimo de revisita. O ciclo termina no momento em que todas as URLs da lista são visitadas. As URLs visitadas são atualizadas no servidor de banco de dados para que entrem na fila aguardando um novo ciclo. Para controlar quais URLs foram visitadas e indexadas, o servidor de banco de dados mantém dados sobre a data, horário e status (informando se a visita foi bem sucedida) da visita para cada URL registrada. A Figura 3-8 ilustra o fluxo de tarefas desempenhadas pelo rastejador.

Para executar a cópia das páginas visitadas, o rastejador utiliza classes que retornam um conjunto de bytes dos dados que serão copiados. Esses dados são convertidos para caracteres ASCII e, ao chegar ao destino, o conteúdo da página é reconstruído de forma que o módulo geo-colaborador possa executar o processo de análise sintática e extração das informações.

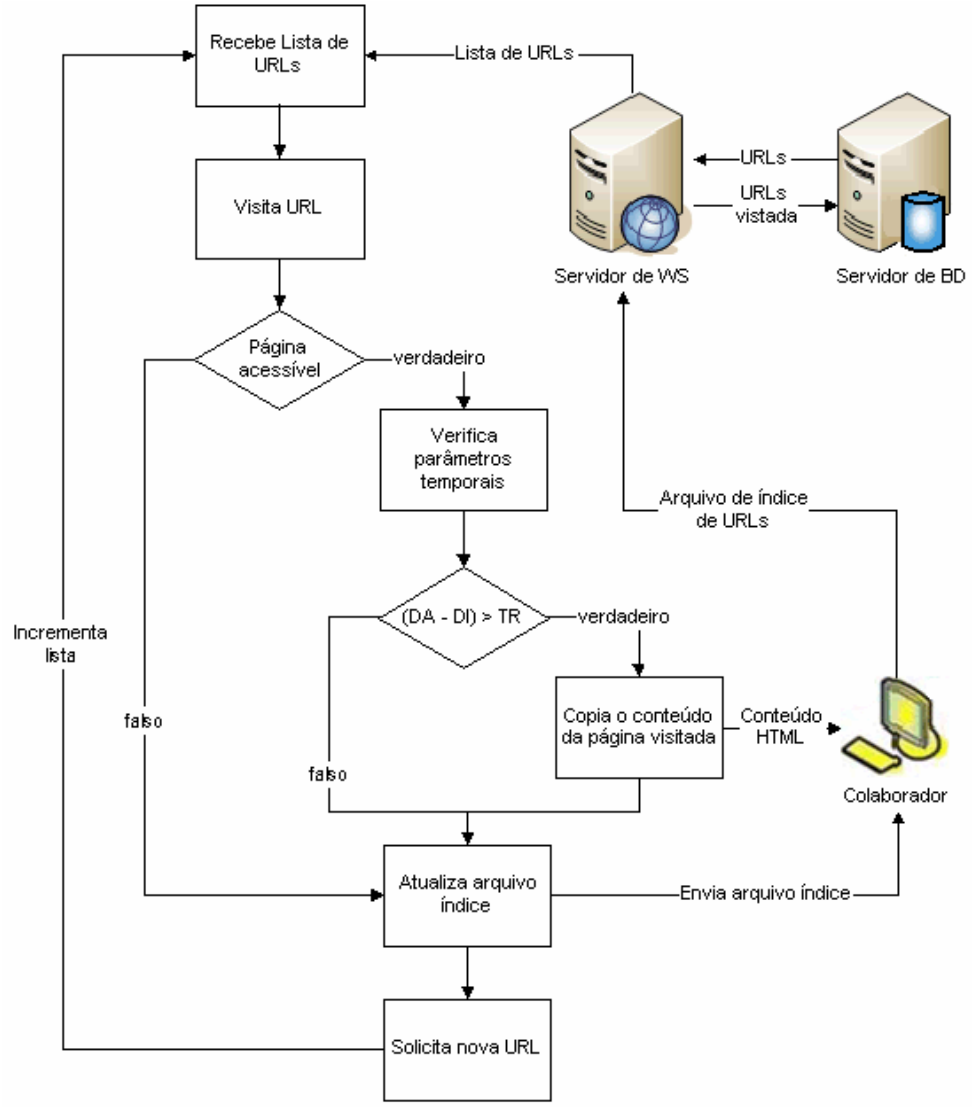


Figura 3-8 – Fluxo tarefas realizadas pelo rastejador.

Os rastejadores utilizados pelo *GeoDiscover* foram desenvolvidos respeitando o "código de ética dos rastejadores" composto por um conjunto de políticas pré-estabelecidas, quais sejam:

- *Política de seleção* que declara qual página deve ser visitada;
- *Política de revisita* que declara quando uma página deve ser revisitada para verificar se houve atualizações;
- *Política de cortesia* que declara como evitar sobrecargas em sítios da web.

A *política de seleção* é garantida através da classificação das URLs catalogadas. Nessa classificação são priorizadas as URLs retiradas de páginas geo-interessantes. A *política de revisita* está diretamente relacionada com a política de cortesia. Para cada acesso a uma página, são armazenadas a data e o horário de acesso. Esses dados são utilizados para reordenar a lista de URLs de forma que a página não seja visitada sucessivamente no mesmo dia. A revisita acontece toda a vez que a lista de URLs é totalmente percorrida. O tempo necessário para que a lista seja percorrida depende da quantidade de URLs que estão indexadas e o número de usuários colaboradores que estão auxiliando no processamento.

A *política de cortesia* é garantida por meio da análise dos protocolos de exclusão de robôs e da *metatag robots*. Toda a vez que uma página visitada explicita regras para não indexação, o *GeoDiscover* descarta seu conteúdo e não executa a indexação. Porém a URL referente a esta página entra na lista de URLs para ser visitada posteriormente. Adotamos esta estratégia, pois o responsável pela manutenção da página em questão pode, posteriormente, modificar as regras, com o objetivo de possibilitar a indexação da página.

No protótipo atual, cada rastejador tem a capacidade de visitar aproximadamente 1 página a cada 10 segundos. Este número é razoável, uma vez que o mecanismo pode envolver centenas de usuários colaboradores aumentando consideravelmente a cobertura da *Web*.

3.6 Análise sintática

O processo de análise sintática do conteúdo HTML das páginas é executado no módulo geo-colaborador e tem como principal função encontrar indícios que possam caracterizar a página como geo-interessante, extrair dados relevantes para a indexação da página e encontrar arquivos geográficos. A Figura 3-9 explicita as tarefas executadas durante a análise sintática dos conteúdos HTML.

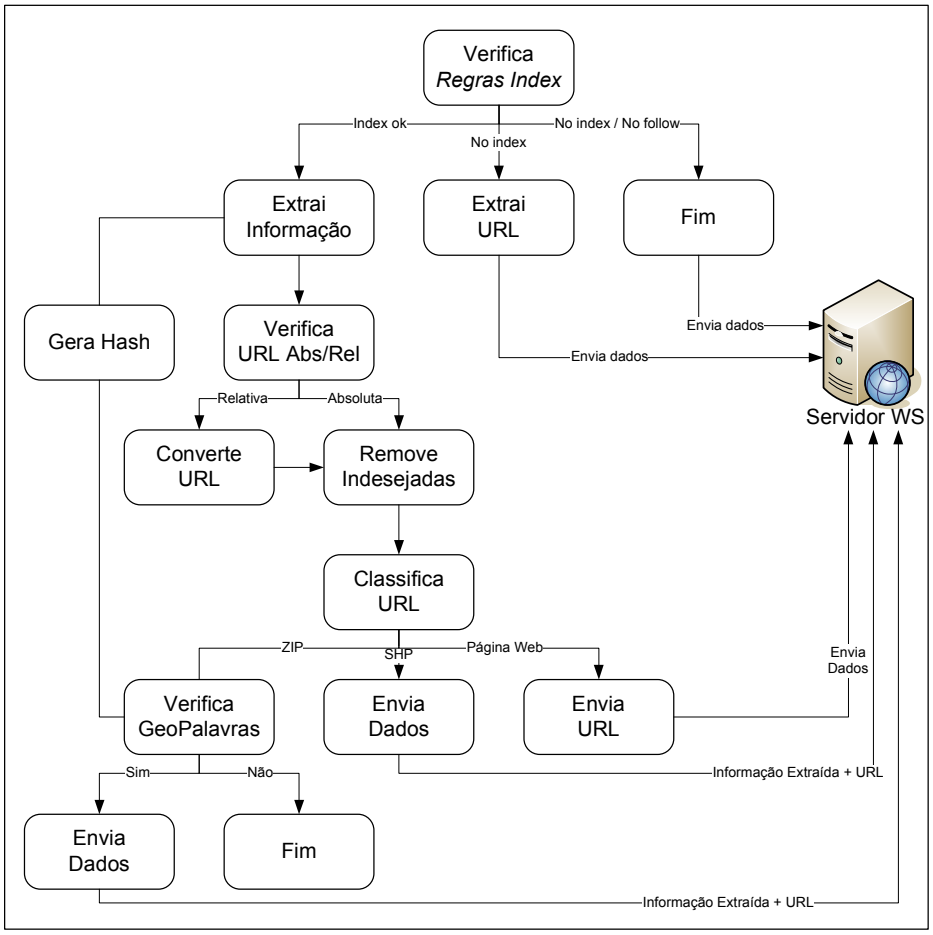


Figura 3-9 – Fluxo tarefas realizadas durante a análise sintática.

Inicialmente são verificadas as regras de indexação da página em respeito ao código de ética dos robôs[10]. Regras para a indexação de páginas foram discutidas em Koster (1995). O *GeoDiscover* foi implementado para verificar duas formas de exclusão de robôs: o arquivo *robots.txt* e a meta-tag especial *robots.*

O protocolo de exclusão de robôs é um padrão que explicita as regras de indexação em um arquivo estruturado denominado *robots.txt* que fica armazenado no servidor *web* da página em questão, em um caminho pré-estabelecido ("/robots.txt"). No início do arquivo *robots.txt* existe um campo denominado `User-agent` que indica o nome dos robôs que deverão respeitar as regras descritas. Pode-se denominar um ou vários robôs específicos ou utilizar `*` para todos os robôs. O próximo campo, `Disallow`, especifica URLs parciais que não

[10] Informações sobre o arquivo *robots* estão disponíveis em http://www.robotstxt.org/wc/norobots.HTML.

poderão ser visitadas. Abaixo é apresentada seqüência de exemplos de utilização do arquivo *robots.txt*.

- Todos os robôs não podem visitar as URLs que iniciam com "/geoinformacao/" ou "/spring/portugues/".

```
# /robots.txt para o site http://www.dpi.inpe.br/
User-agent: *
Disallow: /geoinformacao/
Disallow: /spring/portugues/
```

- Todos os robôs, com exceção do *GeoDiscover*, não podem visitar as URLs que iniciam com "/arquivos/shape/".

```
User-agent: *
Disallow: /arquivos/shape/

User-agent: Geodiscover
Disallow:
```

- Todos os robôs não podem visitar o sítio inteiro.

```
User-agent: *
Disallow: /geoinformacao/
Disallow: /spring/portugues/
```

A *metatag* especial *Robots* é incluída no cabeçalho de uma página HTML e explicita as regras de indexação da página por meio de diretivas separadas por vírgula. As diretivas são `[NO]INDEX e [NO]FOLLOW`, que especificam se uma página pode ou não ser indexada e se as URLs contidas na páginas podem ou não serem seguidas. O valor `robots` pode ser alterado pelo nome de um mecanismo específico, como por exemplo, *google* ou *yahoo*. Qualquer arranjo dos valores *index/noindex, follow/nofollow* é permitido. Tomemos os exemplos abaixo.

- A página pode ser indexada e as URLs constantes na página podem ser utilizadas para serem seguidas pelo robô.

```
<HTML>
<head>
```

```
<meta name=”robots” content=”index,follow”>
<meta name="Descrição" content="Esta página ....">
<title>...</title>
</head>
<body>
```

- A página não pode ser indexada e as URLs constantes na página não podem ser utilizadas para serem seguidas pelo robô.

```
<meta name=”robots” content=”noindex,nofollow”>
```

Em linhas gerais definem-se duas possibilidades distintas quanto às regras de indexação: permissão ou não para indexação completa da página. Porém em alguns casos onde não há permissão para a indexação, existem regras explícitas que permitem que o rastejador visite as URLs presentes na página.

Caso a regra não autorize a indexação da página nem a seqüência dos *links*, o geo-colaborador envia para o servidor WS a confirmação da visita e encerra o processo. Caso a regra não autorize a indexação, mas autorize a seqüência dos links, o geo-colaborador extrai todas as URLs da página e as envia para o servidor WS. Estas URLs serão armazenadas hierarquicamente no BD para serem visitadas posteriormente. Caso a página esteja disponível para indexação, é iniciada a extração do conteúdo das tags *title*, *description*, *keywords*, *href*, *src* e de todas as palavras do corpo da página.

Antes de executar as análises sobre as URLs presentes na página, URLs relativas são convertidas em URLs absolutas. Essa conversão é necessária, pois URLs relativas não fornecem o caminho completo do recurso que descreve (p.e. /*shapefiles.HTML*). Desta forma, para que o rastejador possa executar a visita, é necessário informar as URLs absolutas que contêm o caminho completo do recurso (p.e. *www.mediacenter.com/shapefiles.HTML)*.

Para encontrar arquivos específicos, é realizada uma análise das URLs a fim de identificar arquivos com extensões *shp* e *zip*. Também é executada uma análise sobre a composição do nome dos arquivos com extensões *zip* para detectar a presença das seqüências de letras *shape* e *shp*. Para direcionar o rastejador, evitar processamento desnecessário e sobrecarga de tráfego na rede, foi implementado um algoritmo para a remoção de URLs indesejadas.

As URLs indesejadas são caracterizadas pelas extensões pdf, doc, txt, css, ico, gif, jpg. O processo de remoção de URLs indesejadas é fundamental para o mecanismo, uma vez que evita que geo-colaborador envie para o BD URLs desnecessárias para verificação, diminuindo significativamente a quantidade de URLs indexadas. A URL analisada é classificada em quatro diferentes categorias. Cada categoria exige do geo-colaborador uma ação específica:

- A URL é uma *referência para uma página web* (p.e. http://www.dpi.inpe.br). Neste caso, o geo-colaborador irá enviar para o servidor WS a URL que será armazenada no BD para ser visitada posteriormente.

- A URL é uma *referência para um arquivo shp* (p.e. http://www.ibge.com.br/imagens/saopaulo.shp). Neste caso, o geo-colaborador irá enviar para o servidor WS a URL que será utilizada pelo servidor de aplicações para copiar o arquivo e proceder às etapas de extração de metadados e indexação do mesmo. Também serão enviadas partes do conteúdo HTML para serem armazenadas no servidor de BD e utilizadas no processo de consulta do usuário.

- A URL é uma *referência para um arquivo compactado* (p.e. http://www.census.gov/geo/img/florida.zip). Neste caso, o geo-colaborador irá fazer uma análise para encontrar palavras geo-interessantes no contexto da página. O pressuposto é que páginas que contêm palavras geo-interessantes em seu contexto são fortes candidatas a terem arquivos geográficos. Para a análise, todas as *tags* HTML, *scripts* e outras marcas são removidas, restando apenas texto puro. As palavras existentes neste texto são comparadas com palavras pré-definidas que podem trazer indícios de contexto geográfico. Caso encontre palavras geo-interessantes, o geo-colaborador envia a URL e o conteúdo HTML ao servidor WS. As URLs serão utilizadas pelo servidor de aplicações que irá copiar o arquivo e executar as etapas de descompactação e análise do arquivo. Caso não encontre palavras geo-interessantes o processo é finalizado.

- A URL é uma *referência para um arquivo compactado* e apresenta na composição do caminho ou nome do arquivo as seqüências de caracteres *shape* ou *shp* (p.e.

http://www.census.gov/geo/img/vt02_shp.zip). Neste caso, o geo-colaborador irá proceder da forma quando encontra um arquivo *shp*, descrita anteriormente.

Para obter maior desempenho durante a busca por contexto geográfico, para todas as *strings* é gerado um código *hash*. O código *hash* é utilizado para fazer a comparação das palavras presentes na página com uma lista de código *hash* de palavras que foi enviada pelo servidor de WS juntamente com a lista de URLs para serem visitadas.

3.7 Repositório de palavras geo-interessantes

Conforme discutido anteriormente, partimos do pressuposto de que páginas que contêm palavras geo-interessantes em seu contexto são fortes candidatas a terem arquivos geográficos. Desta forma, criamos um repositório de palavras geo-interessantes que possam denotar indícios de contexto geográficos durante o processo de análise sintática executado nas páginas capturadas pelo rastejador. As palavras do repositório são comparadas às palavras que constam nas páginas analisadas e, quando há ocorrência de igualdade, a página é classificada como geo-interessante.

O repositório de palavras geo-interessantes foi criado a partir da análise de ocorrência de termos em sítios de produtores de dados geográficos. Foram analisados 50 sítios de produtores de dados na língua portuguesa e inglesa. Utilizando o analisador sintático do *GeoDiscover*, identificamos os termos que mais ocorriam nesses sítios. Observamos o local de ocorrência do termo na página (corpo, título e descrição) e formatações especiais nas fontes (tamanho e negrito). A identificação dos termos é possível por meio das *tags* HTML *body, title, description, bold e font.*

Para cada um dos itens observados foram estabelecidos pesos, sendo 1 (um) para cada aparição do termo no corpo da página, 2 (dois) para aparições no título e na descrição da página e 1,25 (um e vinte e cinco) para aparições com tamanho de fonte maior em relação ao corpo da página ou negrito. O critério para a escolha dos pesos está relacionado à importância dos itens analisados na descrição do conteúdo da página. Normalmente palavras constantes no título, na descrição e em destaque no corpo da página, expressam de forma mais precisa o seu conteúdo.

Antes do cálculo de relevância do termo, aplicamos a técnica de *stopwords* e removemos palavras com pouca informação semântica, dentre elas preposições, pronomes e conjunções. Desenvolvemos um algoritmo para auxiliar no cálculo de um fator R para cada termo resultante na lista de termos. Com base nos pesos descritos anteriormente chegamos à expressão:

$$R = NAT + 2ATT + 2{,}5ATD + 1{,}25ATF \text{ onde:}$$

NAT = número de aparições do termo no corpo das páginas.

ATT = número de aparições do termo na *tag title* das páginas.

ADT = número de aparições do termo na *tag description* das páginas.

AFT = número de aparições do termo com destaque na página, caracterizados por fontes destacadas em negrito e com fontes maiores que o restante do documento.

Após calcular o fator *R* para cada termo, encontramos o valor médio de *R* e selecionamos as palavras com valores de *R* superior à média. Na etapa seguinte, desconsideramos palavras cuja semântica não estava relacionada a um "contexto geográfico", tais como arquivo, página, sistema e programa, entre outras. As palavras resultantes em português foram traduzidas para o inglês e vice-versa visando ampliar o repositório sem prejuízo à sua finalidade. A Tabela 3-1 apresenta as palavras geo-interessantes resultantes.

Para aumentar a eficiência no processo de comparação de palavras, no protótipo implementado, para cada palavra geo-interessante do repositório foi gerado um código *hash* que fica armazenado juntamente com a palavra no servidor de BD.

Para a validação do repositório, selecionamos aleatoriamente um novo conjunto com 50 sítios, sendo 25 de produtores de dados geográficos e 25 sítios que possuíam arquivos compactados, porém não geográficos. Analisamos as palavras presentes nos sítios em contraposição com as palavras geo-interessantes. Em todos os sítios de produtores de dados analisados, encontramos pelo ao menos uma palavra geo-interessante. Nos demais sítios, apenas um sítio apresentou a palavra geo-interessante *cidade* em seu contexto.

Tabela 3-1 – Termos que compõem o repositório de palavras geo-interessantes

Termo português	**Termo inglês**	**Termo português**	**Termo inglês**
aerofoto	aerial photography	Mapa	map
limite	boundary	Mapeamento	mapping
shape	shapefile	Mosaico	mosaic
cartografia	cartography	Projeção	projection
cartográfico	cartographic	Espacial	spacial
censo	census	Espectral	spectral
cidade	city	Temporal	temporal
município	county	Satélite	satellite
estado	state	Temático	thematic
país	country	Temperatura	temperature
clima	climate	Tempo	weather
coordenada	coordinate	sensoriamento	sensing
desmatamento	deforestation	Remoto	remote
escala	scale	Topográfico	topographic
geografia	geography	Vegetação	vegetation
geográfico	geographic	Imagem	image
geoprocessamento	geoprocessing	Sig	gis
Outros termos			
shp	arcview	Latitude	longitude
landsat	ikonos	Quickbird	spot
lat	long	Queimada	geo

3.8 Indexação e Classificação de Arquivos Geográficos

Os mecanismos de buscas convencionais possuem funcionalidades para o rastreamento, indexação e consulta de sítios *web* baseadas na ocorrência de palavras e em métodos próprios de classificação, tal como o PageRank (Page *et al.*, 1999). Recentemente surgiram mecanismos voltados para um nicho específico, por exemplo, o Google Scholar e o CiteSeer (Giles *et al.*, 1998) que indexam artigos científicos com base em sua estrutura. Nesta seção apresentaremos as estratégias utilizadas no *GeoDiscover* para a indexação e classificação de arquivos geográficos.

3.8.1 Indexação

No *GeoDiscover* existem três processos distintos de indexação. O primeiro, específico para URLs das páginas visitadas pelo usuário colaborador; o segundo, específico para os sítios de produtores de dados; e o terceiro, para os arquivos geográficos.

O processo de indexação de URLs das páginas visitadas utiliza um arquivo com três elementos em sua estrutura: a URL da página, a data da última indexação e o status da indexação. Assim que uma página é visitada, o módulo geo-colaborador armazena as informações no arquivo e, posteriormente o envia para o Servidor de WS.

A Tabela 3-2 apresenta um exemplo do arquivo de indexação de URLs. O atributo *status* indica a situação da indexação da página, sendo: 0 – não indexada, 1 – indexada com sucesso, 2 – re-indexação sem sucesso. O *status* com valor 0 existe quando a URL em questão nunca foi visitada. A *data da última indexação* é utilizada para controlar o período de re-indexação da página. O *indicador* demonstra se a página de onde a URL foi extraída é um produtor de dados (neste caso é atribuído o valor 1), ou não (valor 0). Para proceder a uma re-indexação o rastejador deve certificar-se de que a condição descrita abaixo é verdadeira; caso não seja, deverá atribuir ao *status* o valor 2:

$$(DA - DI) > TR, \text{ onde,}$$

DA – data e horário atual
DI – data e horário da última indexação
TR[11] – tempo mínimo de revisita

Tabela 3-2 – Arquivo de indexação de URLs

URL	Data indexação	Status	Indicador
http://www.cbers.inpe.br/pt/index_pt.htm	15.07.2006 14:02:10	1	1
http://www.mct.gov.br	15.07.2006 14:02:18	1	1
http://www.brasil.gov.br	15.07.2006 14:02:28	1	1
http://www.inpe.br/institucional/historia.php	15.07.2006 14:03:06	1	1
http://www.inpe.br /institucional/missao.php	15.07.2006 14:03:12	1	1
http://www.inpe.br /institucional/instalacoes.php	15.07.2006 14:02:18	1	1
http://www.inpe.br /cri2/cri_nacional.php	15.07.2006 14:02:29	1	1

[11] No *GeoDiscover*, definimos o TR com 86400 segundos.

O processo de indexação de sítios de produtores de dados é executado pelo Servidor de Aplicações. O processo de indexação se inicia após o analisador sintático, executado no geo-colaborador, extrair as informações de cada produtor e enviá-las ao SA. Essas informações incluem título, descrição, palavras-chave e conteúdo da página; URLs que apontam para outros sítios; e URLs que apontam para arquivos geográficos. O processo de indexação relaciona cada produtor de dados às informações que o descreve, às URLs de arquivos geográficos, à data, ao *status* da indexação e ao escore de importância do produtor. Os critérios definidos para o cálculo da importância de um produtor de dados serão detalhados a seguir. A revisita a um sítio produtor de dados, é garantida pela recolocação da URL do produtor no arquivo de URLs, que serão visitadas posteriormente respeitando as regras de re-indexação descritas acima.

O índice dos arquivos geográficos é criado logo após a análise sintática sobre os mesmos. As informações extraídas são armazenadas e para cada arquivo é criado um identificador que o relaciona com a URL do produtor de dados e com as informações do produtor. Também são armazenadas informações sobre a data e horário da indexação. As informações mantidas do arquivo geográfico são: identificador, descrição de lugar, lugar, tamanho do arquivo, tipo do arquivo, data da indexação, nome do arquivo e quantidade de *downloads* executados a partir do *GeoDiscover*.

O *GeoDiscover* não re-indexa arquivos geográficos, pois uma vez que ele não armazena o arquivo propriamente dito, mas sim uma referência para o arquivo, há apenas uma rotina de verificação de *links* quebrados. Quando um determinado *link* para um arquivo geográfico é quebrado, ele é removido da base de dados. A rotina de verificação de *links* é executada pelo módulo geo-colaborador e segue as mesmas regras de visita para páginas.

3.8.2 Organização e controle da lista de URLs

A lista de URLs é ampliada à medida que os rastejadores visitam e indexam novas URLs. Devido à velocidade desempenhada pelos rastejadores e à grande quantidade de *links* indexados, rapidamente a lista de URLs atinge milhares de itens que deverão ser visitados e indexados. Dessa forma, um dos problemas enfrentados pelo *GeoDiscover* é esgotar a lista de URLs e encontrar o número máximo de produtores de dados, ampliando a abrangência do

mecanismo. Conforme discutido em (Kleinberg *et al.*, 2001), por meio da estrutura de *links*, pode-se encontrar comunidades[12] web:

> *"Pages and links are created by users with particular interests, and pages on the same topic tend to cluster into natural "community" structures that exhibit an increased density of links."*

A análise da estrutura da *web* pode conduzir à melhora dos métodos de acesso e compreensão das informações disponíveis e auxiliar na construção de mecanismos de busca mais eficientes e serviços de busca especializados. Baseados na definição de comunidades apresentada por (Kleinberg *et al.*, 2001), podemos aceitar como razoável a idéia de que *links* extraídos de produtores de dados têm maior probabilidade de apontarem para páginas geo-interessantes do que *links* extraídos de páginas comuns.

Desta forma, a fim de maximizar a eficiência dos rastejadores do *GeoDiscover* na busca por páginas geo-interessantes, priorizamos a visita das URLs extraídas de sítios de produtores de dados, colocando-as no topo da lista de URLs que serão visitadas. Esta decisão é essencial devido à extensão e ao rápido crescimento da lista, a qual pode ser observada na Figura abaixo:

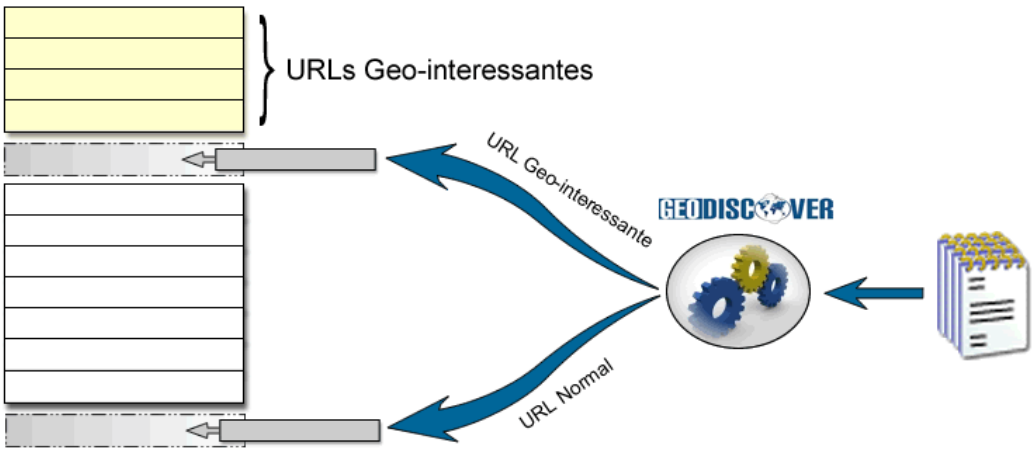


Figura 3-10 – Priorização de URLs.

[12] Uma comunidade pode ser definida como uma coleção de páginas, nas quais cada página membro possui um maior número de *links* apontando para páginas da comunidade do que para páginas que não pertencem à comunidade.

Para tanto, desenvolvemos um **organizador** que inicialmente obtém uma **entrada** com uma lista não ordenada de todas as URLs obtidas pelos rastejadores. A partir dessa lista, uma nova lista é gerada de forma ordenada seguindo os requisitos de prioridade definidos pelo campo *indicador*. Esse campo pode conter dois valores distintos, sendo 1 quando a condição "extraída de um produtor de dados" for verdadeira e 0 caso contrário.

A posição em que a **entrada** será incluída na nova lista é determinada pela sua prioridade, se a entrada for classificada como geo-interessante, será colocada imediatamente após a última **entrada** deste tipo que já se encontra na lista; caso não existam **entradas** deste tipo na lista, ela será colocada no topo; e caso a **entrada** não seja classificada como geo-interessante será adicionada ao final da lista. A Figura 3-11 apresenta o fluxo de tarefas desenvolvidas pelo organizador e a Figura 3-12, uma descrição simplificada, em pseudocódigo, do procedimento de organização de URLs.

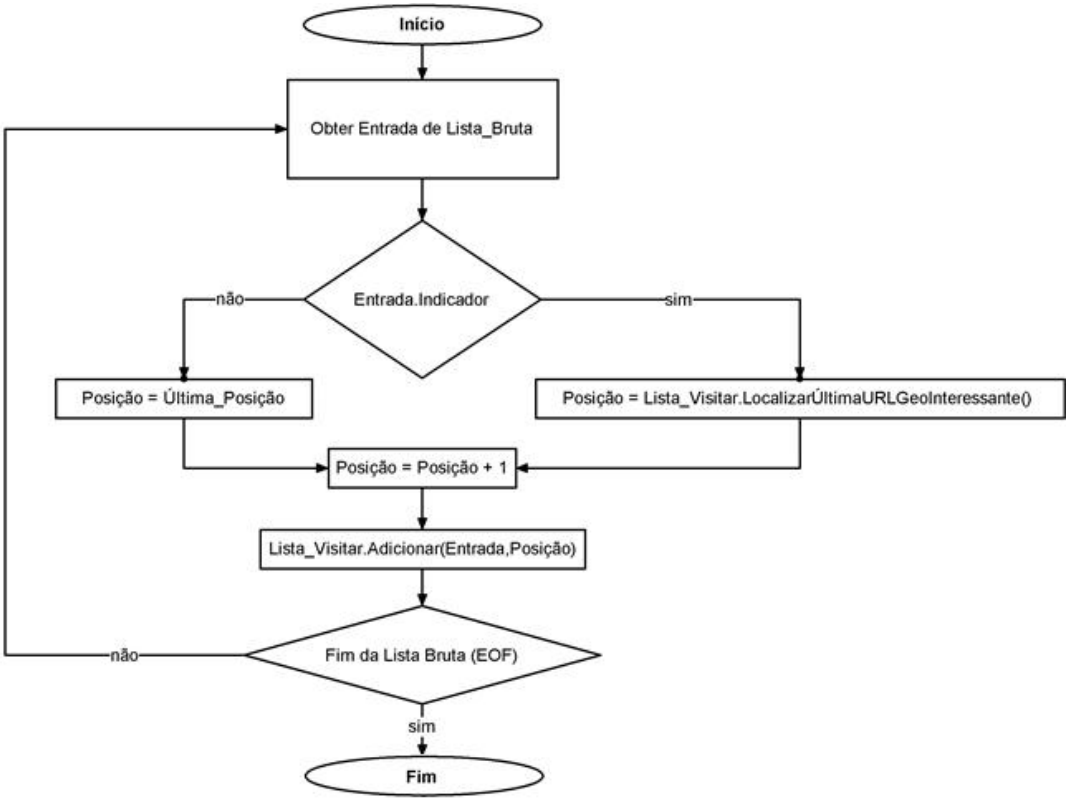


Figura 3-11 – Fluxo de tarefas desenvolvidas pelo organizador de URLs.

Para cada ENTRADA **em** *Lista_Processar* **faça**

Se ENTRADA.indicador = 1 **Então**

Posição_Adicionar = *Lista_Visitar*.**LocalizarÚltimaURLGeoInteressante**() + 1

Lista_Visitar.**Adicionar** ENTRADA **na** *Posição_Adicionar*

Senão

Lista_Visitar.**Adicionar** ENTRADA **na** *ÚltimaPosição*

Fim_Se

Fim_Para

Funções:

`LocalizarÚltimaURLGeoInteressante()` retorna a posição na lista da última **entrada** que foi marcada como geo-interessante. Caso não encontre nenhuma **entrada** desse tipo, retornará o valor 0 (zero).

Figura 3-12 – Algoritmo para organizar e priorizar lista de URLs.

3.8.3 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 essas conexões está sendo pesquisada e utilizada para melhorar a eficiência dos rastejadores (Cho *et al.*, 1998), com a finalidade de auxiliar no processo de classificação de páginas indexadas pelos mecanismos de busca, para descobrir comunidades *web* e para organizar os resultados da pesquisa.

Um *hyperlink* contém a URL para a página a que ele referencia e um texto-âncora que descreve a ligação. O texto-âncora pode oferecer excelentes descrições das páginas a que ele referencia. Estes textos-âncora 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 ele 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). A Figura 3-13 ilustra os conceitos de texto-âncora e texto-âncora estendidos.

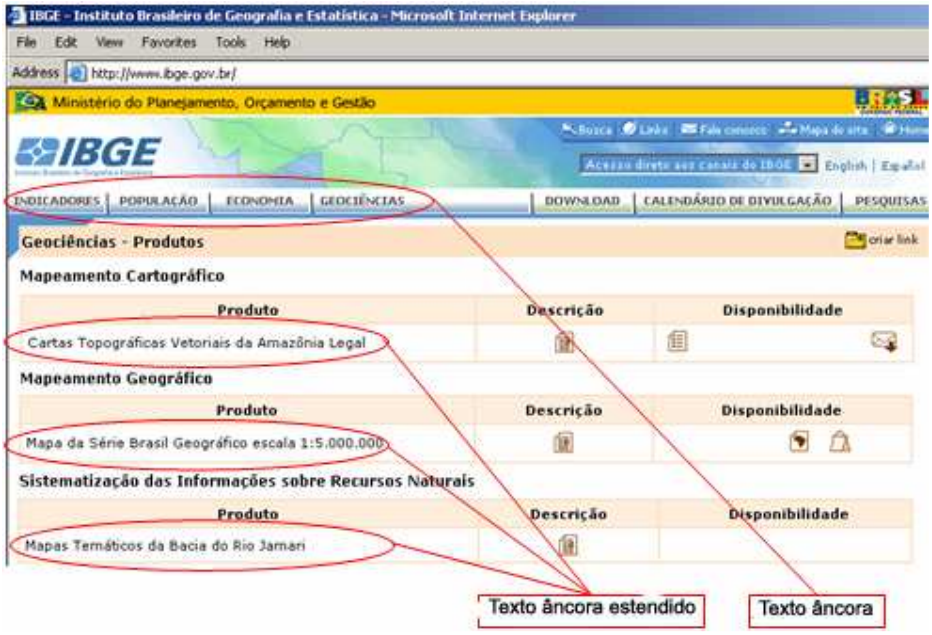


Figura 3-13 – Texto-âncora e texto-âncora estendido.

No *GeoDiscover*, a extração do texto-âncora e do texto-âncora estendido é executada pelo analisador sintático das páginas HTML. Para tanto, são identificadas as *tags href* e seu conteúdo é extraído. No exemplo abaixo, o "arquivo.zip" é extraído, assim como a expressão "mapa da cidade de São José dos Campos na escala 1:20.000"

```
<a href="arquivo.zip">mapa da cidade de São José dos  Campos na escala 1:20.000</a>
```

O *GeoDiscover* utiliza o conceito de texto-âncora em diferentes situações:

- Para auxiliar os rastejadores na localização de arquivos *shape* compactados. Assim como na análise do contexto da página, o texto-âncora é analisado para verificar se contém palavras geo-interessantes. Caso contenha essas palavras e aponte para um arquivo compactado, este será capturado para que o geo-colaborador possa analisá-lo.

- Para incrementar a descrição do contexto. As palavras presentes no texto-âncora e no texto-âncora estendido são armazenadas no servidor de BD. Para a composição do texto-âncora estendido são consideradas 20 palavras antes e 20 palavras depois do *hyperlink*.

- Nos resultados da busca. Para ampliar os resultados obtidos, o *GeoDisocover* utiliza o texto-âncora e o texto-âncora estendido em conjunto com metadados extraídos do arquivo *dbf* associado ao arquivo *shape*.

3.8.4 Construção do repositório de termos descritores de lugares

O processo desempenhado por um mecanismo de busca configura-se em uma "cascata" na qual o rastejador constrói uma coleção que é indexada e passível de ser consultada. Para aumentar a performance do rastejador, o módulo de indexação pode fornecer informações sobre a classificação das páginas *web* para que o rastejador seja mais seletivo e privilegie a visita em páginas mais importantes. A Figura 3-14 demonstra a estrutura em "cascata" de um mecanismo de busca.

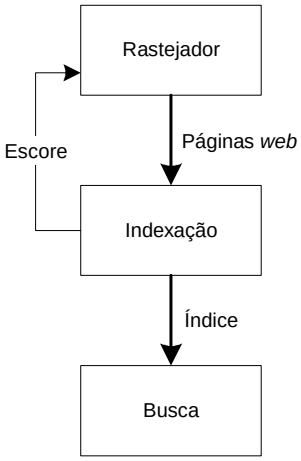


Figura 3-14 – Estrutura em cascata de um mecanismo de busca.

No *GeoDiscover*, a eficiência do processo de busca está relacionada à forma de indexação das coleções de arquivos recuperados. Informações que possam descrever o conteúdo dos arquivos são fundamentais para subsidiar as busca e fornecer resultados relevantes aos usuários.

Uma das formas mais naturais adotadas por um usuário na busca por informações de um determinado local é utilizar o nome (descrição) do local como parâmetro para a consulta em um mecanismo de busca. Um usuário que pretende encontrar arquivos da cidade de São José dos Campos, naturalmente utilizaria a expressão *São José dos Campos* em sua consulta.

O reconhecimento de nomes de lugares é uma condição fundamental para a indexação de recursos *web* (Densham; Reid, 2003). Desta forma, para permitir que o usuário obtenha sucesso na pesquisa de arquivos disponíveis na *web*, é imprescindível que o mecanismo de busca mantenha para cada arquivo geográfico indexado o nome do local que ele representa.

Algumas vezes o nome do local está explícito no texto-âncora do *hyperlink* que aponta para o arquivo, o que facilita a criação de metadados de sua descrição. Nos casos onde não há descrição explícita, é necessário extrair dados adicionais do arquivo.

Para incrementar as descrições dos arquivos geográficos indexados pelo *GeoDiscover* inicialmente desenvolvemos um analisador sintático que verifica a estrutura do arquivo *dbf* associado ao arquivo *shp* a fim de construir um conjunto de ***referência de nomes de lugares*** a partir da identificação de padrões nas nomenclaturas utilizadas nos rótulos das colunas do arquivo *dbf*. O primeiro passo para o desenvolvimento do analisador foi estabelecer um repositório de **termos de descritores de locais**. Para isto foram analisados 100 arquivos *dbf* associados a arquivos *shp* obtidos aleatoriamente de diversos sítios produtores de dados.

Na amostra coletada, procuramos diferenciar as fontes dos arquivos para evitar a predominância de padrões individuais. Através da amostra, pôde-se verificar inicialmente que os nomes dos campos dos bancos de dados associados são boas referências para se procurar uma informação desejada. A partir da análise, observou-se que há uma tendência dos nomes dos campos descreverem, de forma direta ou indireta, o conteúdo dos mesmos. Por exemplo, ocorrências de nomes de campos tais como “Cidade” e “*Location*”, servem como bons descritores de seu conteúdo e esses campos, por sua vez, caracterizam o nome do local representado pelo *shape*.

A partir do número de aparições de nomenclaturas de campos nos arquivos *dbf* foi possível identificar uma coleção de **termos descritores de locais**. A fim de ampliar as possibilidades de nomenclaturas, sobre os termos selecionados, aplicamos a técnica de *stemming* para

extrair a origem morfológica das palavras, utilizando o algoritmo Porter Stemming (Porter, 2006). A coleção resultante de radicais originou um repositório de **termos descritores de lugares** que é utilizado para a comparação executada pelo analisador sintático. O principal benefício obtido com a utilização de radicais é a ampliação das possibilidades de comparação, que é executada considerando possíveis prefixos e sufixo.

Para a alimentação do ***repositório de nomes de lugares*** dos arquivos *shape* indexados, desenvolvemos um algoritmo que percorre os nomes dos campos contidos no arquivo *dbf* associados ao arquivo *shape*, compara os nomes encontrados com o repositório de descritores de lugares previamente estabelecido, identifica os campos que descrevem nomes de lugares e cria uma lista invertida destes campos. Para cada item da lista invertida é gerado um código *hash*, que será utilizado pelo mecanismo para responder às consultas dos usuários. A Figura 3-15 ilustra duas tabelas de arquivos *dbf* e a lista invertida resultante.

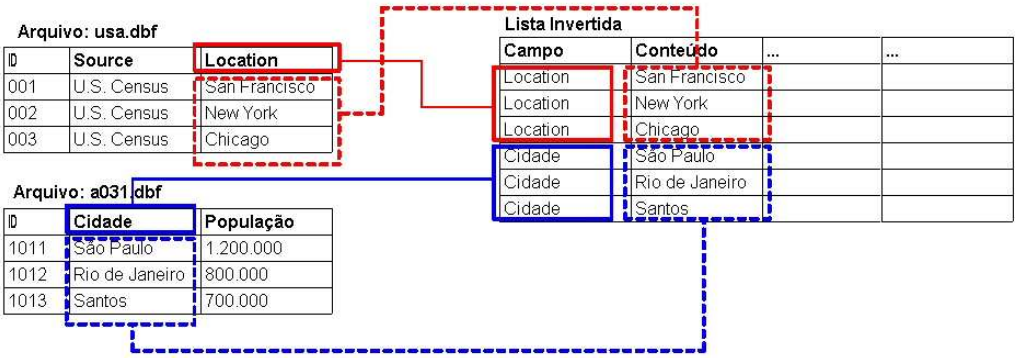


Figura 3-15 – Lista invertida resultante de arquivos *dbf*.

Porém, após a execução de testes sobre novos arquivos *dbf* foi possível verificar que as nomenclaturas de descritores de lugares nos rótulos dos campos variam muito e, geralmente não expressam o significado do que descrevem. Nomenclaturas como, por exemplo, *nn01* e *cty1996* foram encontradas e descreviam nomes de cidades e nomes de países respectivamente.

Frente às limitações impostas pela metodologia inicialmente empregada e a fim de ampliar o repositório de descritores de lugares, fundamental para o processo de recuperação de arquivos geográficos, implementamos uma função para comparar o conteúdo dos campos

dos arquivos *dbf* a um *gazetteer*. Utilizamos o *gazetteer* da Alexandria Digital Library (ADL, 1999) que possui aproximadamente 6 milhões de ***nomes de lugares***.

Uma coluna de banco de dados descreve um atributo de um mesmo domínio e para cada atributo há um conjunto de valores permitidos (Silberschatz *et al.*, 2002); para o atributo *nome_município*, por exemplo, o domínio é o conjunto de todos os nomes de municípios. Com base nessa definição, implementamos um algoritmo que utiliza apenas os 5 primeiros valores da coluna a ser analisada. Esta decisão diminui significativamente o tempo de processamento durante a comparação. A Figura 3-16 apresenta as etapas para a construção de termos descritores de locais com a utilização do *gazetteer*.

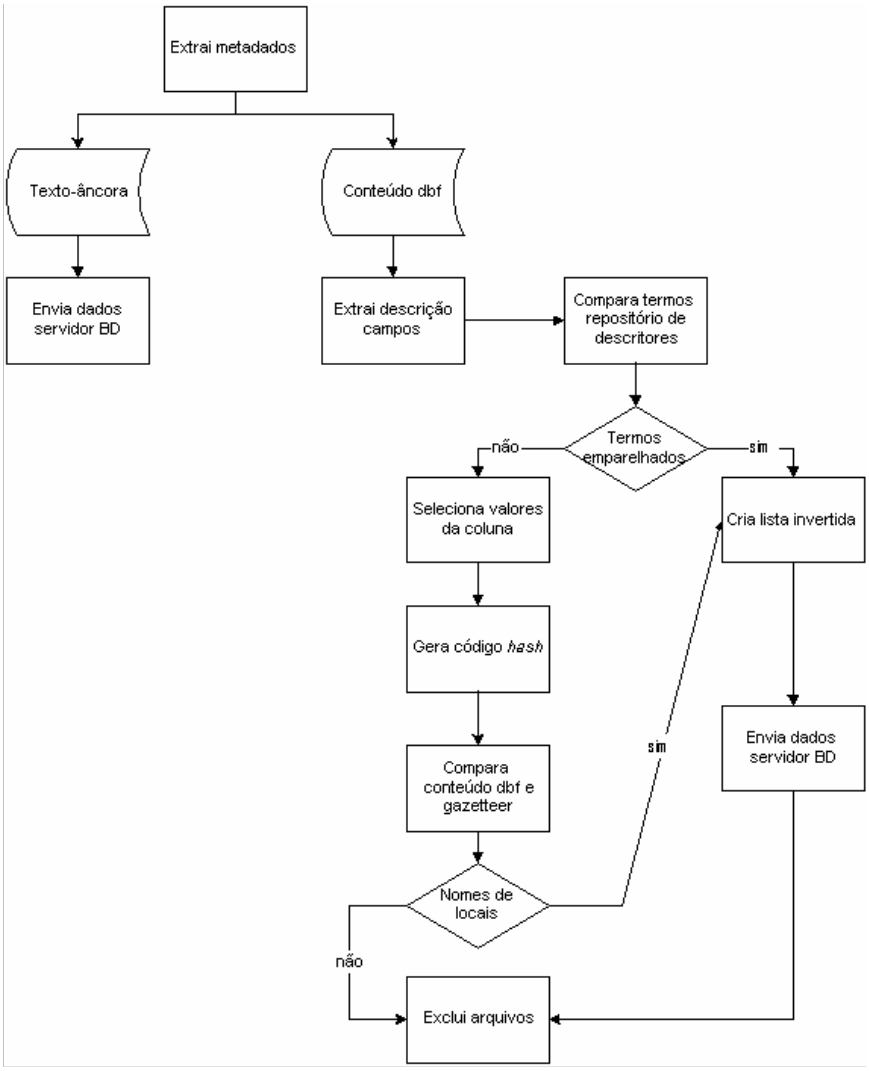


Figura 3-16 – Processo para a construção de termos descritores de locais

Inicialmente são extraídos metadados do texto-âncora do *hyperlink* que aponta para o arquivo geográfico. Esses metadados são armazenados no servidor de BD e são utilizados para responder às consultas dos usuários. O analisador sintático geográfico examina a estrutura do arquivo *dbf* e extrai os rótulos (descrição) dos campos que ele contém. Cada campo é comparado com os **termos descritores de lugares**. Caso o rótulo da coluna combine com algum termo descritor, a descrição e os valores da coluna são incluídos na lista invertida e armazenados no servidor BD. Caso o rótulo não combine com nenhum termo, os 5 primeiros valores de cada coluna são comparados com os ***nomes de lugares*** contidos no *gazetteer*. Para ganho de performance durante a execução da comparação, para cada palavra do *gazetteer* foi gerado um código *hash* que está armazenado no servidor de BD e para cada valor da coluna é gerado um código *hash* antes da comparação. Desta forma, comparamos o código *hash* das duas coleções. Caso os valores combinem, a descrição e os valores da coluna em questão são incluídos na lista invertida e armazenados no servidor de BD. Finalizado o processo de comparação, os arquivos *dbf* são excluídos.

3.8.5 Considerações sobre a sobreposição de técnicas para a extração de metadados em arquivos geográficos

Para a composição de metadados dos arquivos *shape* indexados pelo *GeoDiscover*, e conseqüentemente possibilitar aos usuários consultas mais precisas, utilizamos importantes fontes de dados: o conteúdo do arquivo dbf, o texto-âncora do *link* que aponta para o arquivo geográfico e o nome do arquivo.

A metodologia empregada combinando os descritores de locais e o *gazetteer* amplia drasticamente a probabilidade de identificação do nome do lugar. Assim, criamos a possibilidade do usuário pesquisar diretamente pelo nome do lugar para a obtenção de uma resposta mais próxima de seu interesse durante a execução da consulta. A utilização do texto-âncora permite extrair dados importantes para a descrição do arquivo geográfico, como o nome do lugar, o tipo do dado (mapa, carta topográfica, rede de pontos) e a escala que podem possibilitar o refinamento da consulta.

O nome do arquivo normalmente define uma “classe” para os dados que ele representa. Como exemplo, podemos citar o arquivo “rios.shp/ rios.dbf”, que contém registros

informativos sobre rios. Um determinado campo desse arquivo com o rótulo “Nome” provavelmente conterá valores com os nomes dos rios presentes no arquivo; por outro lado, um campo com o rótulo “Nome” em um arquivo “cidades.shp / cidades.dbf” conterá valores diferentes do anterior. Essa dedução é conseqüência de observações de resultados de pesquisas executadas durante o desenvolvimento desta tese. Com isso, podemos considerar adequada, para fins de pesquisa específica, a utilização do nome dos campos como parâmetro para saber “onde” procurar no arquivo. Exemplificando, há uma diferença entre uma pesquisa do tipo “cidades de São Paulo” e outra do tipo “São José dos Campos”. No primeiro caso não se trata de uma busca por conteúdo específico, mas sim uma busca por “classe” (qualquer cidade do Estado de São Paulo satisfaria a consulta), enquanto que a segunda contempla um conteúdo específico (uma única cidade).

É importante ressaltar que os arquivos shape normalmente são disponibilizados em pacotes compactados para facilitar sua distribuição, uma vez que um *shape* é composto por 3 ou mais arquivos (*.shp, *.shx e *.dbf). Houve casos em que nestes pacotes estava incluído um arquivo descritor do conteúdo apresentado, bem como dos campos do banco de dados associado, estabelecendo uma fonte de metadados. Em todos os casos observados, esses arquivos possuíam o mesmo nome do arquivo shape associado à extensão texto (*.txt), conforme exemplificado na Figura 3-17. A análise sobre este arquivo pode ser incluída durante o processo de indexação, com o objetivo de se obter mais informações descritivas do arquivo shape visando melhorar os resultados retornados a uma consulta.

```
==================================
GISMAPS - www.gismaps.com.br
==================================
Mapas Temáticos - Geoprocessamento
==================================

* Grupo: América do Norte

* Titulo: Principais rodovias da América do Norte

* Descrição: Mapa apresentando as principais rodovias dos EUA

* Ano/Revisão: 2000.2

* Data de atualização: 21/06/2005

* Fonte: Intergraph

* Conteúdo:

********** rodovias1.dbf **********
ROUTE_NUMB      -número da rota/estrada
ROUTE_NUM       -número da rota/estrada secundário

==================================
GISMAPS - www.gismaps.com.br
----------------------------------|
```

Figura 3-17 – Arquivo metadados de um arquivo *shape*.

CAPÍTULO 4

IMPLEMENTAÇÃO E AVALIAÇÃO DO *GEODISCOVER*

Neste capítulo apresentaremos detalhes de implementação e as interfaces utilizadas no *GeoDiscover* para a consulta e visualização de arquivos geográficos e para o acompanhamento das tarefas executadas pelo módulo geo-colaborador. As estratégias adotadas para a ordenação dos resultados utilizando a classificação de produtores de dados também são discutidas neste capítulo. Finalmente, apresentaremos os resultados obtidos em testes executados sobre o protótipo implementado.

4.1 Protótipo implementado

O mecanismo de busca, denominado GeoDiscover, foi implementado em C# utilizando recursos do *framework* .NET 2.0, e para o gerenciamento do banco de dados utilizamos o SGBD relacional SQL Server 2000. Foram desenvolvidos componentes que abrangem as funções necessárias para o processo completo de busca, descoberta, captura, indexação, consulta e recuperação de arquivos geográficos: rastejamento da web (algoritmo especializado em arquivos geográficos), análise sintática de contexto (páginas *HTML*), análise sintática de arquivos geográficos (*shape* e *dbf*) e interface para consulta aos dados indexados e interface . Para o processo de descompactação dos arquivos utilizamos as classes nativas `FileInputStream`, `ZipInputStream` e `ZipEntry`, que estão contidas no *namespace* java.util.zip, desenvolvido em J# e referenciado pelo C# por meio da biblioteca de vínculo dinâmico *vjslib.dll*. Para a manipulação de arquivos *shape*, utilizamos a biblioteca de ligação dinâmica (*dll*) de código aberto SharpMap[13]. A Figura 4-1 apresenta a interação dos processos, os fluxos de dados e os componentes do protótipo implementado.

[13] A biblioteca pode ser acessada no sítio http://sharpmap.iter.dk/.

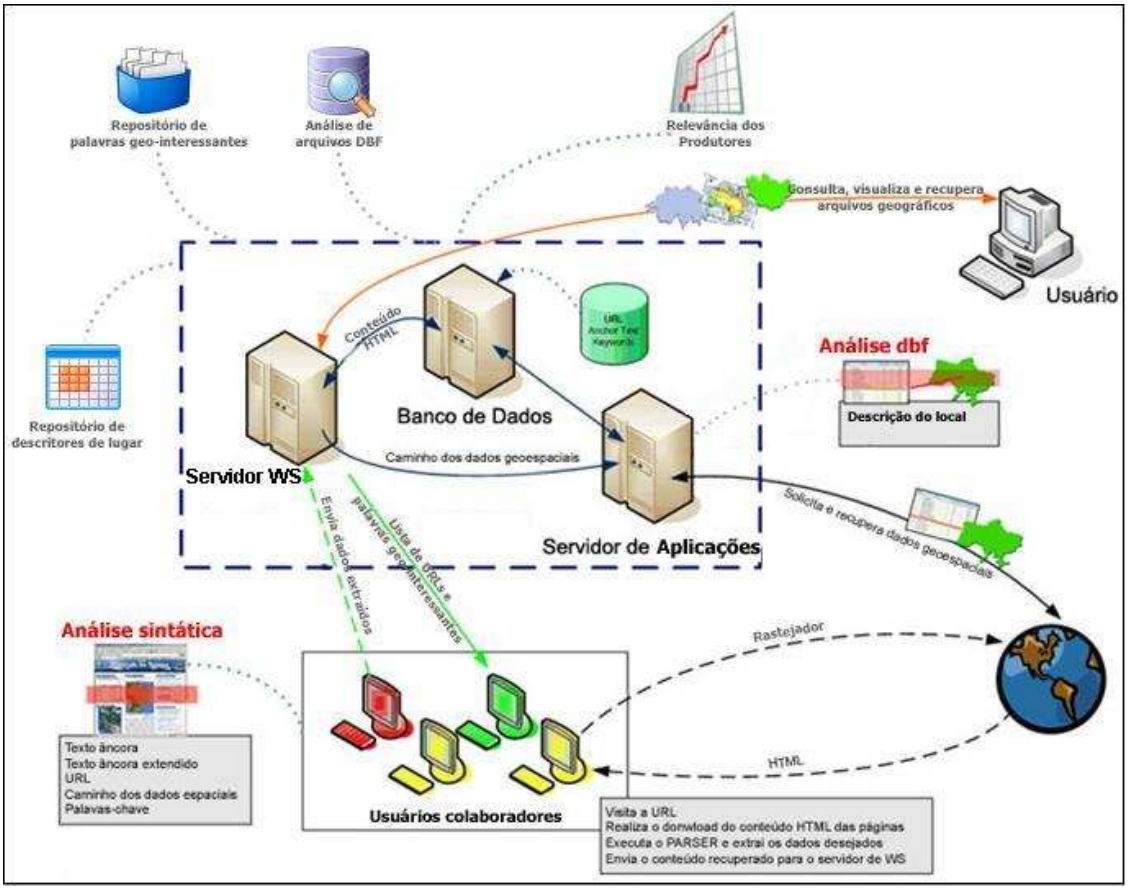


Figura 4-1 – Fluxo de dados e componentes do *GeoDiscover*.

4.2 Consulta e visualização

Conforme apresentado anteriormente, o *GeoDiscover* permite a recuperação de arquivos geográficos (*shape*) por meio da análise integrada do contexto da página e de metadados extraídos do próprio arquivo. A interface de consulta foi desenvolvida para possibilitar que usuários leigos executem consultas de maneira simples e intuitiva por meio de nomes de lugares. Também desenvolvemos uma interface para que os arquivos recuperados sejam visualizados adequadamente. As interfaces foram construídas para serem visualizadas em navegadores padrão e podem ser acessadas remotamente pela Internet no endereço *www.geodiscover.org*.

Estrategicamente definimos duas possibilidades de consulta: direta ou ampliada. A consulta direta faz a comparação dos termos fornecidos pelo usuário com o repositório de nomes de

lugares e com o texto-âncora do *link* que aponta para o arquivo geográfico. Dessa forma, os resultados obtidos tendem a ser mais precisos. A consulta ampliada, além de utilizar o repositório de nomes de lugares e o texto-âncora, utiliza também o texto-âncora estendido e o contexto do sítio do produtor de dados. Desta forma, alcançamos maior abrangência dos resultados, porém perdemos em precisão. A figura abaixo exemplifica os metadados envolvidos nas consultas diretas e ampliadas:

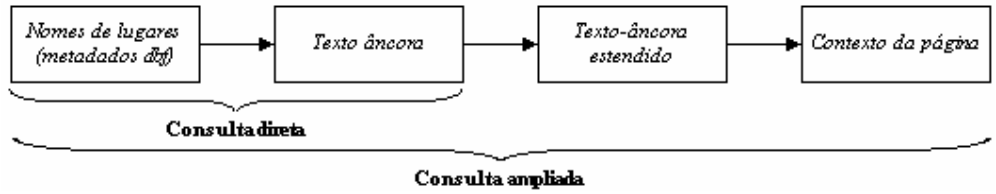


Figura 4-2 – Metadados envolvidos nas consultas diretas e ampliadas.

Conforme demonstrado na Figura 4-3, a interface de consulta é amigável, e permite que o usuário digite o nome de lugar e proceda a consulta direta, utilizando o botão *Search* ou ampliada, por meio do *hyperlink Expanded Search*.

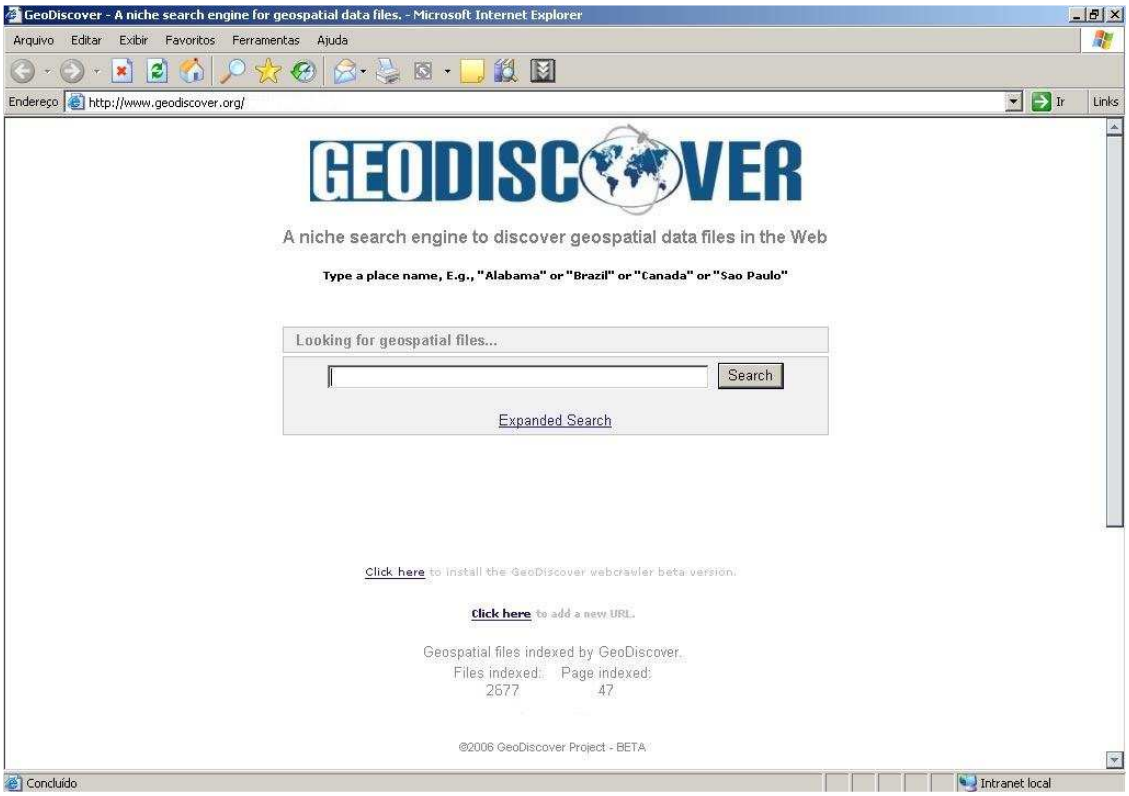


Figura 4-3 – Interface de consulta do *GeoDiscover*.

Os nomes de lugares fornecidos pelo usuário são convertidos para código *hash* e então são comparados com os metadados correspondentes de cada arquivo geográfico indexado, de acordo com as regras definidas para as consultas diretas e ampliadas. A comunicação entre o usuário e o Servidor de Aplicações, responsável pelo processamento das consultas, é intermediada pelo Servidor de *Web Services*. Os metadados dos arquivos que satisfazem os critérios da consulta são ordenados seguindo as regras de classificação dos produtores de dados e são disponibilizados para o usuário em uma interface que combina informações do arquivo e informações do produtor de dados. Caso não encontre nenhum arquivo que atenda à consulta, retorna para o usuário uma mensagem de advertência. A Figura 4-4 apresenta as tarefas desenvolvidas durante o processo de consulta.

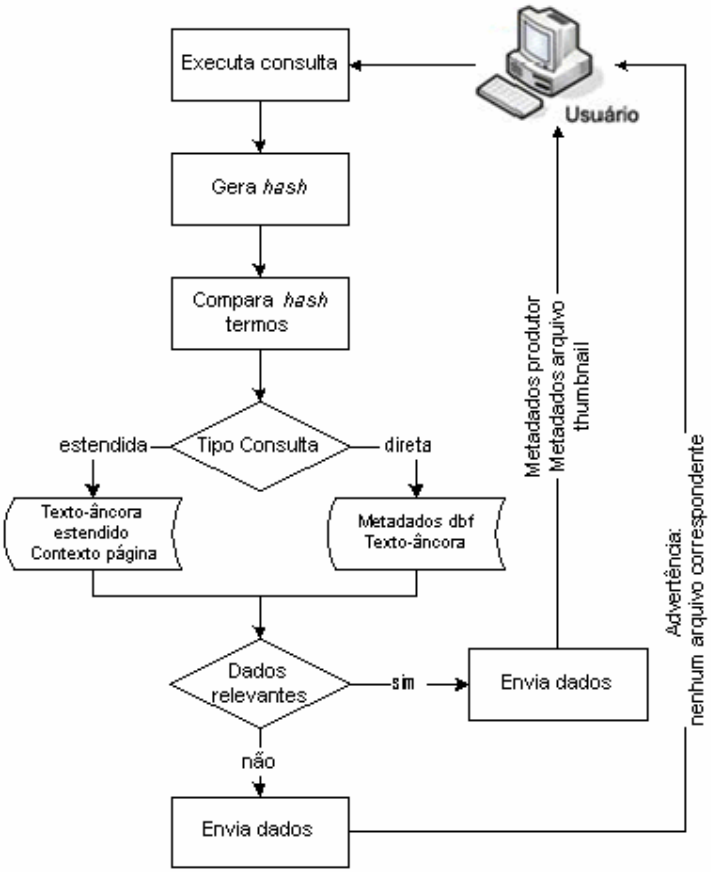


Figura 4-4 – Fluxo de tarefas envolvidas na consulta.

Conforme demonstrado na Figura 4-5, para cada item presente na lista de arquivos é apresentado, no quadro à direita, um *thumbnail* e informações adicionais do arquivo, tais como nome, tipo e quantidade de *downloads* executados a partir do *GeoDiscover*. Também

são disponibilizadas informações sobre o produtor de dados, tais como sua URL, título e descrição do sítio, e a quantidade de dados geográficos disponíveis no sítio deste produtor. Informações adicionais sobre o arquivo podem ser obtidas clicando sobre o *thumbnail* ou sobre o *hyperlink* "*More Details*". Estas informações são convenientes e proporcionam aumento de produtividade, uma vez que auxiliam o usuário a conhecer melhor os dados geográficos que são adequados às suas necessidades antes de iniciar o processo de *download*. A organização das interfaces é baseada na convicção de que a consulta por nomes de locais e a visualização dos resultados com imagens são formas mais naturais, intuitivas e preferidas pelos usuários.

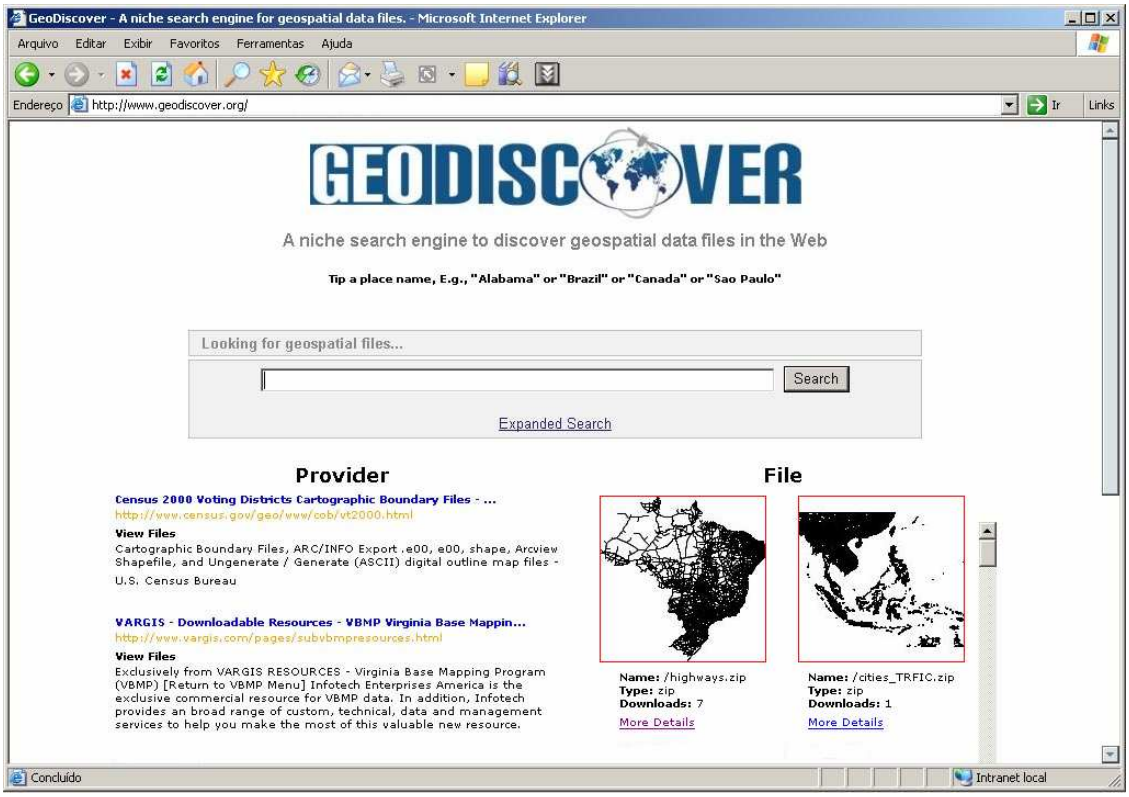


Figura 4-5 – Interface de apresentação de resultados.

4.3 Interface do usuário colaborador

Conforme discutido anteriormente, para se tornar um colaborador, o usuário necessita instalar em seu computador um módulo gerenciador, denominado geo-colaborador. Para que o usuário possa monitorar os dados processados em seu computador pelo geo-colaborador,

desenvolvemos uma interface gráfica que está apresentada na Figura 4-6. Essa interface permite que o usuário inicie ou termine a execução do geo-colaborador e disponibiliza as seguintes informações:

- ***Tags* extraídas**, como o **título** (p.e. *<title>DPI - Divisão de Processamento de Imagens</title>*), descrição (p.e. *<meta name="description" content="A Divisão de Processamento de Imagens (DPI) faz parte da Coordenação Geral de Observação da Terra (OBT) do Instituto Nacional de Pesquisas Espaciais (INPE) ">* e **palavras chave** (p.e. <*meta name="keywords" content*=" Ciência da Geoinformação, Geoprocessamento, GIS, SIG, Bancos de Dados Geográficos">.

- **URLs localizadas**, absolutas e relativas que o colaborador enviará para o servidor WS. São extraídos os conteúdos das *metatags href* e *src* (p.e. <a href="http://www.census.gov/geo/www/cob/metadata.HTML"> e <src="http://www.census.gov/geo/www/cob/new_buttons/images/button.gif">).

- **Texto-âncora e texto-âncora estendido,** a partir da posição da palavra em relação à URL a que o texto está próximo (p.e. *Country shape files, World, also includes some rivers and populated places*).

- **Caminho** de arquivos geo-interessantes, que serão armazenados e analisados posteriormente (p.e. <*a href*="/geo/cob/bdy/vt/vt00e00/vt02_d00_e00.zip"> e <a href="misc/SouthAmerica_adm0.shp">).

- **Conteúdo HTML** completo da página, incluindo todas as marcações (p.e. <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN"> <!-- saved from url=(0023)http://www.dpi.inpe.br/ --> <HTML><HEAD><META content="Geoprocessamento, Ciência da Geoinformação, GIS, Sistema de informação geográfica, SIG, Bancos de Dados Geográficos" name=Keywords>.

- **Páginas Geo,** que listam as URLs de páginas que foram classificadas como geo-interessantes.

- ***Link–Server*** que lista todos as URLs recebidas do servidor que serão visitadas pelo colaborador.

- **Palavras** extraídas da página sem a marcação (p.e. DPI, Divisão, Processamento, Imagens, Geoprocessamento, Ciência, Geoinformação, GIS, Sistema, informação, geográfica, SIG, Bancos, Dados, Geográficos).

- ***Status* do serviço,** que demonstra todas as ações executadas pelo rastejador.

- **Log de erros,** que lista todos os erros ocorridos durante a execução da aplicação, como por exemplo, páginas não encontradas e falhas na conexão.

- ***Robots,*** que mostra o conteúdo do arquivo Robots.txt ou o conteúdo da *metatag robots* da página que está sendo analisada.

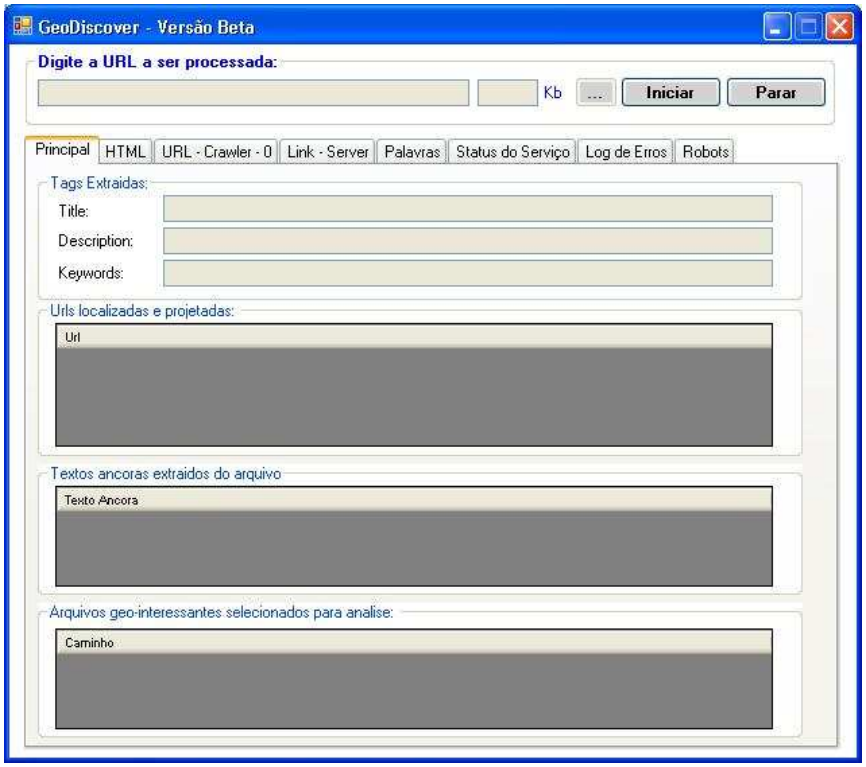


Figura 4-6 – Interface do módulo geo-colaborador.

4.4 Classificação de produtores de dados

Mecanismos de busca tradicionais utilizam diferentes formas para classificar os principais sítios da *Web.* O *Google*, utiliza o método de *PageRank* para priorizar os resultados de busca pela ocorrência de palavras (Page *et al.*, 1999). Similarmente ao método para classificar

páginas utilizando a informação dos links, o mecanismo *Citeseer (Giles et al., 1998)* classifica artigos científicos como *hubs* e autoridades baseado no gráfico de citações.

O *GeoDiscover* ordena os produtores de dados baseado em três aspectos: quantidade de dados geográficos disponíveis no sítio da *Web*; quantidade de *downloads* executados pelos usuários a partir dos resultados apresentados pelo *GeoDiscover*; e pelo escore do produtor, medido pela quantidade e qualidade de apontamentos que ele recebe.

O escore é computado a partir da estrutura de *links* existentes entre as URLs indexadas e pela classificação de um sítio como produtor de dados ou não. Essa classificação é fundamental para a ordenação, pois ela condiciona todos os relacionamentos existentes entre os sítios (*hyperlinks*). Nesta abordagem, definimos um peso *P* para cada apontamento. Caso o apontamento seja feito por um sítio produtor de dados, classificado previamente, o valor de *P* é 2; caso o apontamento seja feito por um sítio convencional (não produtor de dados), o valor de *P* é 1. O escore é o resultado da soma de todos os pesos *P* para um determinado sítio. A Figura 4-7 exemplifica a estrutura de *links* e o valor *P* para cada apontamento.

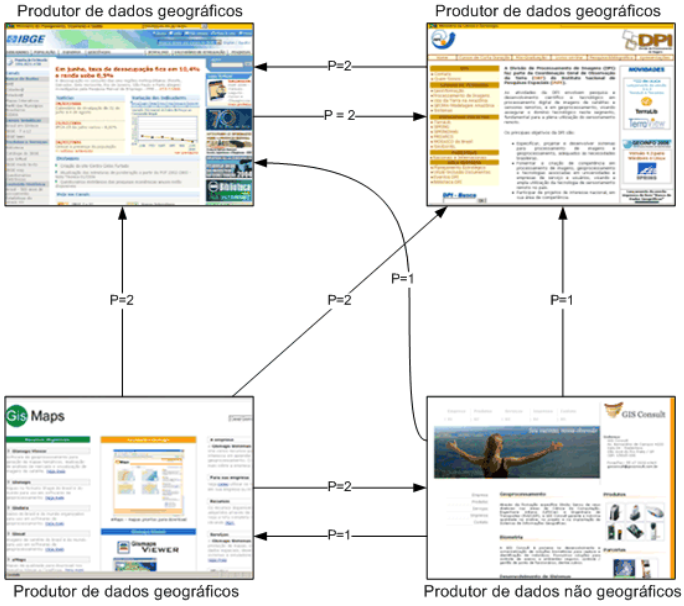


Figura 4-7 – Estrutura de *links* e peso *P* para os apontamentos.

O processo para o cômputo do escore, demonstrado na Figura 4-8, acontece após a extração das URLs de um determinado sítio. Um mapeamento dos *links* é criado indicando as URLs de origem e as URLs de destino. No mapeamento são introduzidas apenas as URLs absolutas, e

auto-apontamentos são desconsiderados. A partir do mapeamento são identificados os apontamentos, o valor do peso *P* e o escore de cada URL é incrementado. A Figura 4-9 faz uma analogia ao mapa de *links* e a Figura 4-10 apresenta o escore das páginas demonstradas no exemplo.

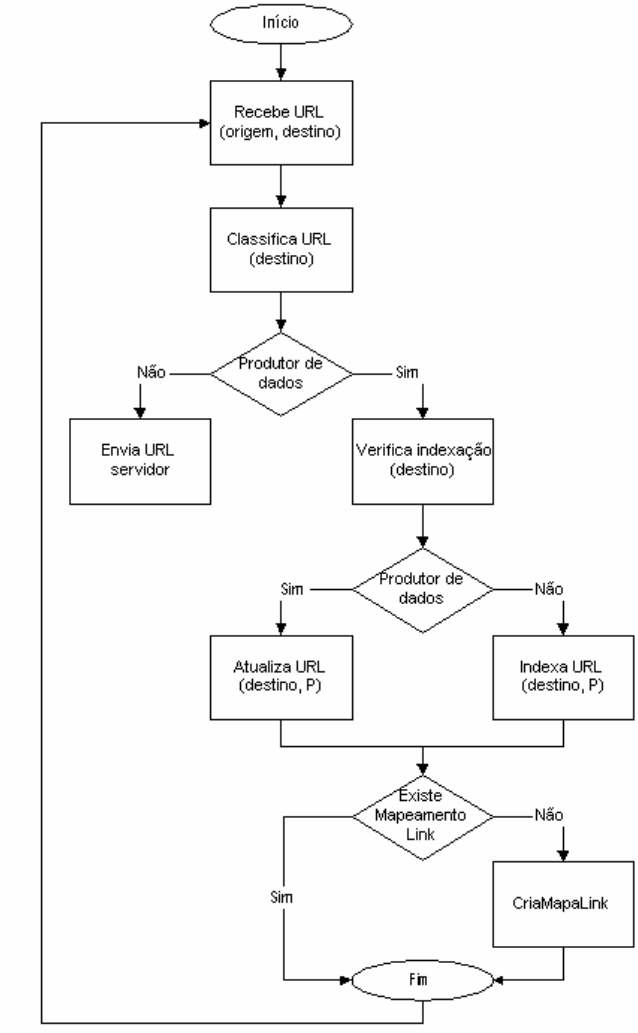


Figura 4-8 – Etapas envolvidas no cômputo do escore.

Desta forma, o escore do produtor de dados é utilizado para ordenar os resultados retornados ao usuário, seguido pela quantidade de arquivos e pela quantidade de *downloads* efetivados. Ao retornar uma lista de arquivos geográficos para o usuário, o *GeoDiscover* apresenta, para cada arquivo, seu produtor (por exemplo, IBGE), a quantidade de arquivos geográficos produzidos por este produtor (por exemplo, 52 arquivos) e a quantidade de *downloads*

executados utilizando a interface do *GeoDiscover* (por exemplo, 5 *downloads*). Esses dados são importantes para que o usuário possa verificar a procedência dos dados retornados, conhecendo, *a priori*, o produtor do dado que ele poderá capturar.

Origem	Destino	Peso (P)
http://www.ibge.gov.br	http://www.inpe.br	2
http://www.inpe.br	http://www.ibge.gov.br	2
http://www.gismaps.com.br/	http://www.inpe.br	2
http://www.gismaps.com.br/	http://www.gisconsult.com.br	2
http://www.gismaps.com.br/	http://www.inpe.br	2
http://www.gisconsult.com.br	http://www.gismaps.com.br/	1
http://www.gisconsult.com.br	http://www.inpe.br	1
http://www.gisconsult.com.br	http://www.ibge.gov.br	1

Figura 4-9 – Mapeamento de *links*.

URL	Soma (P)	Escore
http://www.ibge.gov.br	2 + 1	3
http://www.inpe.br	2 + 2 + 2 + 1	7
http://www.gismaps.com.br/	1	1
http://www.gisconsult.com.br	2	2

Figura 4-10 – Escore das páginas obtido a partir do mapeamento de *link*s.

4.5 Desempenho do *GeoDiscover*

O protótipo desenvolvido apresentou um desempenho satisfatório. O tempo médio necessário para, a partir de um URL, visitar a página, fazer a cópia, e analisá-la é menor que 10 segundos. Dessa forma, um usuário colaborador tem a possibilidade de indexar 360 páginas por hora, o que representa em média 8.640 páginas por dia. Com base nesses números, podemos prever que com 10 usuários colaboradores trabalhando durante um mês, teríamos cerca de 2.5 milhões de páginas indexadas e com 1000 usuários, este número se elevaria para aproximadamente 2.5 bilhões de páginas.

Para fins de testes de eficiência, ajustamos o protótipo atual para indexar 15.000 páginas a partir da URL inicial *www.ibge.gov.br*. Essa URL foi escolhida aleatoriamente dentre outros produtores de dados brasileiros.

A Figura 4-11 demonstra os resultados obtidos. Foram encontrados e indexados 47 produtores de dados. Analisando os resultados, pudemos observar que a proporção de produtores de dados indexados em relação às páginas visitadas é de aproximadamente 1/320. Nestes 47 sítios de produtores de dados, encontramos 2.677 arquivos geográficos com

extensões *.shx, *.exe, *.rar entre outras, porém constatamos que a extensão predominante foi a *.zip. Com relação ao tamanho físico dos arquivos geográficos, a coleção encontrada somou 6.5 Gb na sua totalidade, resultando um tamanho médio de 2.5 Mb por arquivo. Este tamanho médio inclui os arquivos *shape*, *índice* e *dbf* associados ao mesmo objeto. Finalmente verificamos que o tamanho total da base de metadados corresponde a aproximadamente 0,2% do tamanho dos arquivos geo-interessantes que eles descrevem.

Sítios visitados	Produtores de dados	Número de arquivos geo-interessantes	Tamanho total dos arquivos geo-interessantes	Tamanho médio dos arquivos	Tamanho total da base de metadados
15.000	47	2.677	6.5 Gb	2.5 Mb	12 Mb

Figura 4-11 – Resultados obtidos em testes preliminares do *GeoDiscover*.

É importante ressaltar que esses resultados podem variar drasticamente, pois fatores não controlados pelo *GeoDiscover* podem interferir nesses resultados. Dentre esses fatores, estão as características físicas da conexão utilizada pelos colaboradores e a organização e o tamanho das páginas indexadas. Outro fator que merece atenção é o tempo necessário para copiar os arquivos geográficos de sua origem para o Servidor de *Download* do *GeoDiscover*. Trabalhando com uma banda nominal de 512k, foram necessárias 38 horas para o *download* completo de todos os 2.677 arquivos. Nota-se que, nesse caso específico, houve um equilíbrio entre o tempo necessário para a visitação das 15.000 páginas (aproximadamente 41 horas) e o tempo necessário para cópia dos arquivos (aproximadamente 38 horas).

Fica evidente a necessidade de expansão dos servidores de *download* à medida que novos colaboradores aderirem ao projeto, a fim de evitar gargalos no processo de análise dos arquivos geográficos e conseguir manter o equilíbrio do mecanismo.

4.6 Abrangência e precisão

O objetivo de um mecanismo de busca é atender plenamente a demanda de seus usuários, permitindo que eles recuperem informações relevantes às suas necessidades. Entretanto, este objetivo nem sempre é atendido quando a recuperação envolve a *web*. Sabe-se ainda que não há uma padronização universal em relação à relevância. O conceito de relevância é subjetivo, cognitivo, relativo às necessidades atuais de um usuário e pode se alterar ao longo

do tempo. Desta forma, medir a relevância dos resultados de um mecanismo de busca é uma tarefa complexa.

Para avaliarmos a eficiência do protótipo implementado, optamos pela utilização de duas métricas comprovadamente eficientes na avaliação de mecanismos de busca: *abrangência* e *precisão*. A avaliação foi aplicada a duas funcionalidades do *GeoDiscover*: a indexação de arquivos geográficos e a recuperação dos arquivos por meio de uma consulta realizada por um usuário. A *abrangência* é uma medida que avalia matematicamente o quão eficiente é o mecanismo ao retornar resultados relevantes. A *precisão* avalia a exatidão do mecanismo ao retornar estes resultados. As fórmulas adotadas foram:

$$abrangência = \frac{NR_{recuperado}}{NR_{total}} \qquad precisão = \frac{NR_{recuperado}}{N_{recuperado}}$$, onde:

$NR_{recuperado}$	=	Número de resultados relevantes recuperados
NR_{total}	=	Número de documentos relevantes na coleção
$N_{recuperado}$	=	Número total de resultados recuperados

Para o cômputo da *abrangência* e *precisão*, inicialmente preparamos um conjunto de sítios coletados aleatoriamente na *Web*. A partir de uma URL inicial (*www.ibge.gov.br*) indexamos 15.000 sítios na *Web*, sendo 47 sítios de produtores de dados geográficos. Destes 15.000, selecionamos um conjunto com 1.000 sítios, incluindo os 47 sítios de produtores. Após executar a indexação ou a consulta, para cada arquivo relevante na lista ordenada de resultados foram calculadas a *abrangência* e a *precisão* com base em sua posição na lista. Para melhor compreensão dos resultados geramos um gráfico para cada item avaliado.

Para a avaliação da funcionalidade de indexação de arquivos geográficos, instituímos como desafio a recuperação dos sítios de produtores de dados contidos no conjunto. A lista resultante apresentou os 47 sítios de produtores de dados nas 47 posições iniciais, o que atingiu o limiar máximo para *abrangência* e *precisão*, ou seja, a *abrangência* atingiu 100% sem prejudicar a *precisão*, que também se manteve em 100%. Este resultado era esperado pelo fato de estarmos avaliando rastejadores especializados em um determinado tipo de

arquivo, porém cabe ressaltar que estes percentuais demonstram que a indexação é eficiente. A Figura 4-12 apresenta a lista resultante (à esquerda) e os respectivos cálculos para *abrangência* e *precisão* e a Figura 4-13 apresenta o gráfico resultante.

Posição	Relevante		
1	X	→	A = 1/47 = 0,02 P = 1/1 = 1
2	X	→	A = 2/47 = 0,04 P = 2/2 = 1
3	X	→	A = 3/47 = 0,06 P = 3/3 = 1
4	X	→	A = 4/47 = 0,09 P = 4/4 = 1
5	X	→	A = 5/47 = 0,11 P = 5/5 = 1
6	X	→	A = 6/47 = 0,13 P = 6/6 = 1
7	X	→	A = 7/47 = 0,15 P = 7/7 = 1
8	X	→	A = 8/47 = 0,17 P = 8/8 = 1
9	X	→	A = 9/47 = 0,19 P = 9/9 = 1
10	X	→	A = 10/47 = 0,21 P = 10/10 = 1
11	X	→	A = 11/47 = 0,23 P = 11/11 = 1
12	X	→	A = 12/47 = 0.26 P = 12/12 = 1
13	X	→	A = 13/47 = 0,28 P = 13/13 = 1
14	X	→	
15	X	→	A = 15/47 = 0,32 P = 15/15 = 1
16	X	→	A = 16/47 = 0.34 P = 16/16 = 1
17	X	→	A = 17/47 = 0.36 P = 17/17 = 1
18	X	→	A = 18/47 = 0,38 P = 18/18 = 1
19	X	→	A = 19/47 = 0,40 P = 19/19 = 1
20	X	→	A = 20/47 = 0.43 P = 20/20 = 1
21	X	→	A = 21/47 = 0.45 P = 21/21 = 1
22	X	→	A = 22/47 = 0.47 P = 22/22 = 1
23	X	→	A = 23/47 = 0,49 P = 23/23 = 1
24	X	→	A = 24/47 = 0.51 P = 24/24 = 1
25	X	→	A = 25/47 = 0,53 P = 25/25 = 1
26	X	→	A = 26/47 = 0.55 P = 26/26 = 1
27	X	→	A = 27/47 = 0,57 P = 27/27 = 1
28	X	→	A = 28/47 = 0,60 P = 28/28 = 1
29	X	→	A = 29/47 = 0.62 P = 29/29 = 1

30	X	→	A = 30/47 = 0,64 P = 30/30 = 1
31	X	→	A = 31/47 = 0,66 P = 31/31 = 1
32	X	→	A = 32/47 = 0,68 P = 32/32 = 1
33	X	→	A = 33/47 = 0,70 P = 33/33 = 1
34	X	→	A = 34/47 = 0,72 P = 34/34 = 1
35	X	→	A = 35/47 = 0,74 P = 35/35 = 1
36	X	→	A = 36/47 = 0,77 P = 36/36 = 1
37	X	→	A = 37/47 = 0,79 P = 37/37 = 1
38	X	→	A = 38/47 = 0,81 P = 38/38 = 1
39	X	→	A = 39/47 = 0,83 P = 39/39 = 1
40	X	→	A = 40/47 = 0,85 P = 40/40 = 1
41	X	→	A = 41/47 = 0,87 P = 41/41 = 1
42	X	→	A = 42/47 = 0,89 P = 42/42 = 1
43	X	→	A = 43/47 = 0,91 P = 43/43 = 1
44	X	→	A = 44/47 = 0,94 P = 44/44 = 1
45	X	→	A = 45/47 = 0,96 P = 45/45 = 1
46	X	→	A = 46/47 = 0,98 P = 46/46 = 1
47	X	→	A = 47/47 = 1,00 P = 47/47 = 1

Figura 4-12 – Lista resultante e respectivos cálculos para *abrangência* e *precisão*.

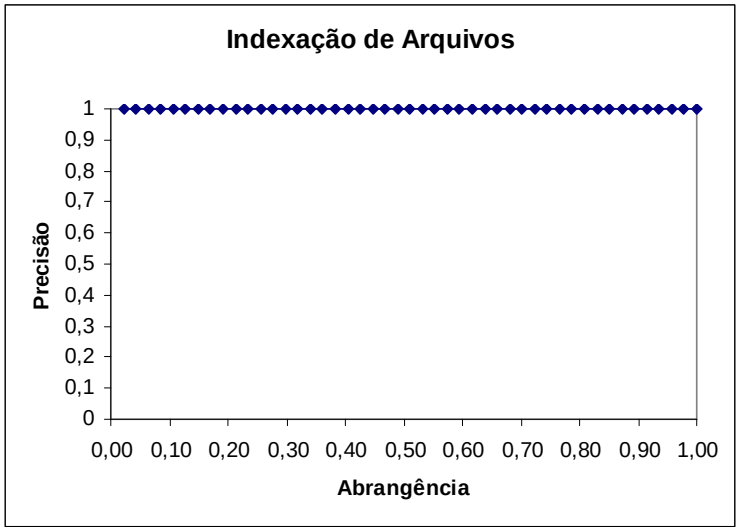


Figura 4-13 – *Abrangência* e *precisão* para o processo de indexação.

Para a avaliação da funcionalidade de recuperação dos arquivos por um usuário, instituímos como desafio a recuperação de todos os arquivos cujo nome de lugar fosse “são paulo”. Avaliamos as duas possibilidades de consulta disponíveis no *GeoDiscover*: direta e ampliada. Para a consulta direta, foram retornados 6 arquivos relevantes até a 9ª posição, conforme pode ser observado na Figura 4-14. O gráfico resultante é apresentado na Figura 4-15.

Posição	Relevante	
1	X	A = 1/6 = 0,17 P = 1/1 = 1
2	X	A = 2/6 = 0,33 P = 2/2 = 1
3	X	A = 3/6 = 0,50 P = 3/3 = 1
4		
5	X	A = 4/6 = 0.67 P = 4/5 = 0.80
6		
7	X	A = 5/6 = 0.83 P = 5/7 = 0.71
8		
9	X	

Figura 4-14 – Posição dos documentos relevantes retornados na consulta direta.

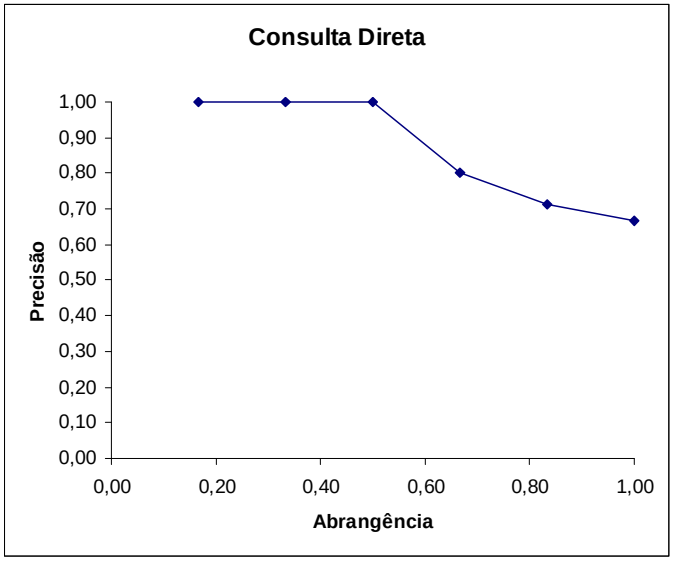


Figura 4-15 – Gráfico de *abrangência* e *precisão* para o processo de consulta direta.

Para a consulta ampliada, executamos o mesmo procedimento de avaliação, porém foram retornados 6 arquivos relevantes até a 28ª posição, conforme pode ser observado na Figura 4-16 e o respectivo gráfico na Figura 4-17.

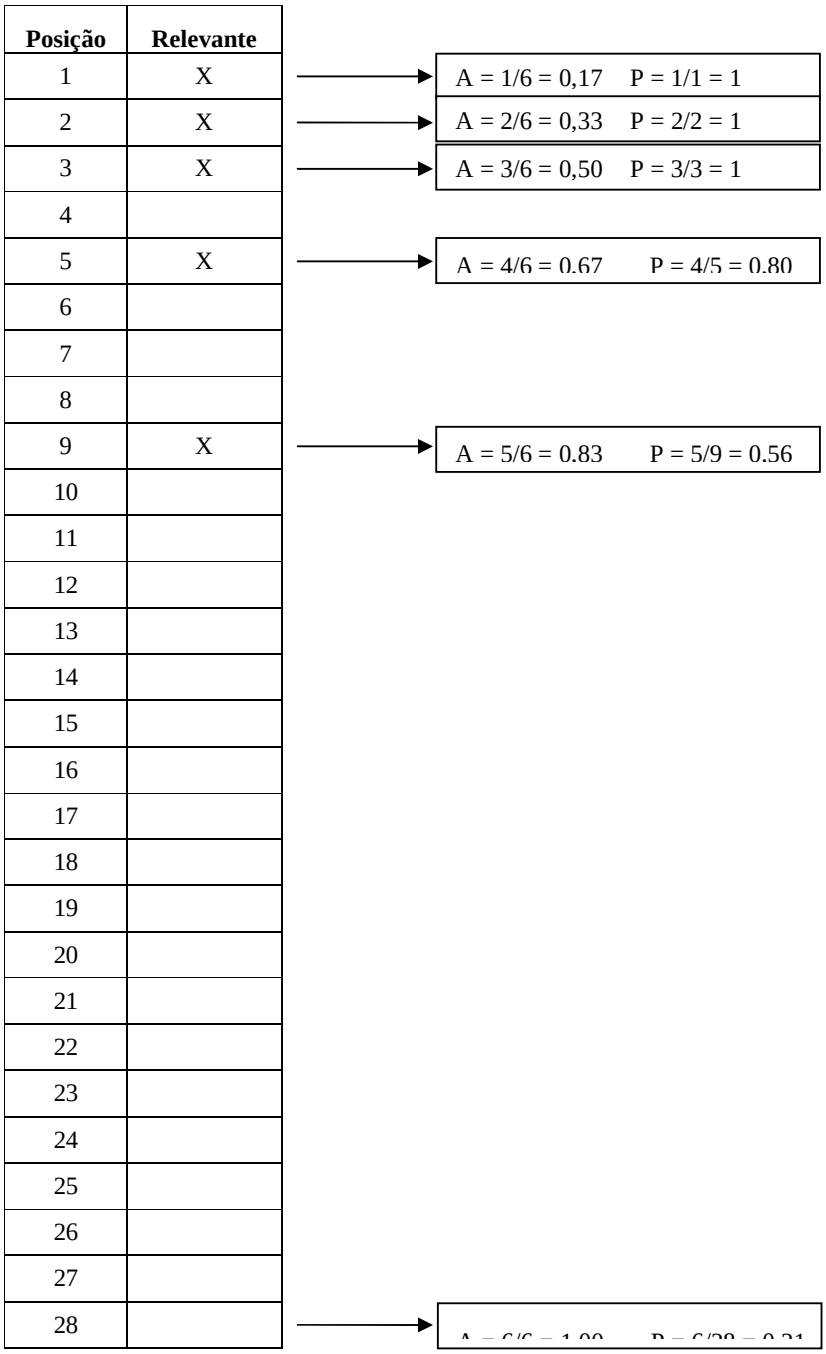


Figura 4-16 – Posição dos documentos relevantes retornados na consulta ampliada.

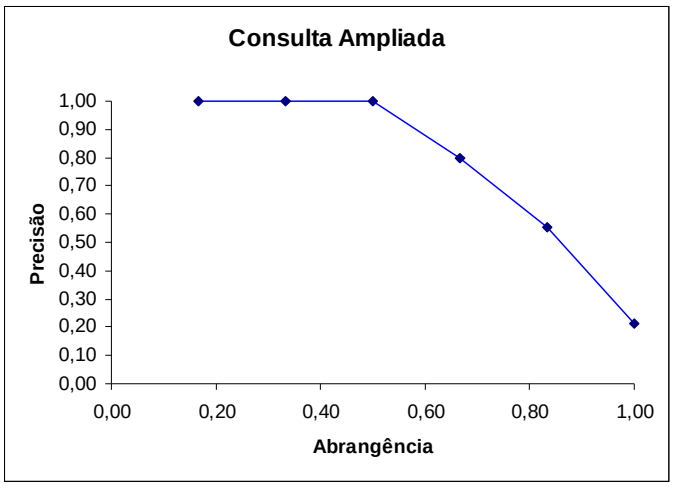


Figura 4-17 – Gráfico de *abrangência* e *precisão* para o processo de consulta ampliada.

Ao analisar os resultados obtidos para os processos de consulta direta e ampliada, concluímos que os resultados da consulta direta foram satisfatórios, pois, como é conhecido, à medida que a abrangência se aproxima de 100%, a precisão tende para 0. Apesar desse declínio ter sido observado na consulta direta, a precisão se estabeleceu em aproximadamente 67%, o que pode ser considerado um bom resultado.

Os resultados obtidos pela consulta ampliada demonstraram desempenho inferior ao observado na consulta direta, pois ao alcançar 100% de abrangência, o número de arquivos não relevantes cresceu significativamente, levando a precisão próxima dos 21%. Porém não podemos ignorar a utilidade da consulta ampliada, que permitirá ao usuário obter uma quantidade maior de respostas para sua consulta. Essa opção é útil quando o usuário não está procurando um arquivo de um lugar específico ou quando tem dúvidas sobre a denominação do lugar.

CAPÍTULO 5

CONCLUSÕES E TRABALHOS FUTUROS

Esta tese apresenta como contribuição principal a metodologia para a descoberta e captura de arquivos geográficos a partir do contexto da página em busca de indícios de conteúdo geográfico com a utilização de rastejadores especializados em arquivos *shape* e *zip*. Além disso, aponta novas técnicas para a indexação e classificação dos arquivos utilizando a análise combinada de metadados extraídos do arquivo geográfico e do sítio do produtor de dados.

Os anos 90 marcaram um crescimento plausível da *Web*, transformando-a num dos mais difundidos meios de disseminação de informação. Paralelamente ao crescimento da *Web*, ocorreu a efetivação da utilização de sistemas de informações geográficas como ferramenta de apoio à decisão e planejamento em diferentes âmbitos da gestão pública e privada. Dessa forma, a necessidade por informação geográfica amplia-se proporcionalmente à quantidade de informações geográficas geradas e disponibilizadas na *Web* por produtores de dados.

A limitação imposta pelos mecanismos de busca tradicionais a conteúdos especializados, particularmente arquivos geográficos, estabelece uma lacuna entre as informações e os usuários que dela necessitam. Para minimizar essa lacuna, desenvolvemos um mecanismo de busca especializado, denominado *GeoDiscover*, com funcionalidades de descoberta, indexação, classificação e consulta a arquivos geográficos.

A utilização de usuários colaboradores para o processo de rastejamento e análise sintática de páginas web distribuído, traz benefícios significativos quanto à possibilidade de cobertura da web, redução de investimentos para a aquisição de servidores poderosos e excelente capacidade de expansão, de maneira sustentável, à medida que novos usuários façam parte do projeto. A interface disponibilizada para o usuário colaborador acompanhar o processo de rastejamento e indexação é interessante no sentido de permitir que o colaborador tenha acesso às informações processadas em seu computador. Porém o processo de gerenciamento

da distribuição de URLs é uma tarefa complexa que deve ser rigorosamente controlada a fim de evitar desvios no processo de rastejamento.

A busca explícita por arquivos *shape*, a partir da extensão *shp*, inicialmente implementada, mostrou-se ineficiente devido à grande quantidade de arquivos compactados disponíveis nos sítios dos produtores. Para superar esse problema, especializamos os rastejadores em arquivos compactados com a extensão *zip*[14] e utilizamos o contexto dos sítios dos produtores para filtrar os arquivos, concentrando o processo de análise sintática dos arquivos somente em páginas classificadas como geo-interessantes. Com a adoção dessa estratégia, reduzimos consideravelmente a quantidade de arquivos compactados que são copiados para os servidores do *GeoDiscover* para análise e aumentamos a quantidade de arquivos *shape* indexados.

Ainda no que diz respeito à redução de arquivos indesejáveis para rastejamento e análise sintática, outro benefício direto ao desempenho do mecanismo foi a remoção de URLs indesejadas caracterizadas pelas extensões pdf, doc, txt, entre outras, da lista de URLs que é enviada ao Servidor de BD com endereços para serem visitados.

A criação do repositório de palavras geo-interessantes mostrou ser de suma importância para a classificação das páginas como geo-interessantes ou não. A utilização dos termos descritores de lugar para a análise do conteúdo dos arquivos *dbf* foi limitada. Porém a combinação desse procedimento com a utilização de *gazetteers* para a caracterização do lugar obteve excelente precisão nos resultados alcançados.

No que diz respeito à recuperação de dados com o *GeoDiscover*, o uso de técnicas para a descrição do nome do lugar aliadas ao texto-âncora, à relevância dos produtores de dados e à interface amigável do mecanismo permite que usuários leigos encontrem arquivos de interesse. A interface do mecanismo é de fácil utilização e a apresentação dos metadados do produtor de dados com metadados e *thumbnail* (que permite uma pré-visualização do arquivo que será copiado) do arquivo *shape* proporciona ao usuário maior facilidade e velocidade durante a filtragem dos arquivos que satisfaçam suas necessidades dentre os resultados retornados a uma consulta.

[14] A escolha pelo formato *zip* baseou-se na predominância deste formato junto aos produtores analisados.

A hipótese de trabalho adotada para a tese foi comprovada, pois já que, baseado nos conceitos adotados, houve o desenvolvimento de um mecanismo de busca especializado em arquivos geográficos. Os resultados obtidos por meio de experimentos realizados permitiram avaliar a metodologia proposta, bem como os componentes do protótipo implementado. Os testes preliminares de eficiência, executados sobre o protótipo, apresentaram um equilíbrio entre abrangência e precisão e demonstraram que o *GeoDiscover* tem capacidade para indexar arquivos geográficos em grande escala, sem prejudicar os servidores *Web* que visita.

Trabalhos futuros relacionados a esta tese devem ser direcionados à utilização de ontologias para nomes de lugares e relacionamentos espaciais a fim de retornar informações mais precisas às consultas executadas pelos usuários. Desta forma seria possível a recuperação de dados a partir de consultas mais genéricas (por exemplo, todas as cidades contidas no estado de São Paulo, ou as cidades vizinhas ao Vale do Paraíba).

Uma das limitações do mecanismo proposto é a possibilidade de recuperar apenas arquivos *shape*; desta forma, há necessidade de ampliação dos rastejadores a fim de reconhecerem outros formatos de arquivos geográficos, tais como arquivos *spr* e *geotiff*. Essa ampliação é viável, uma vez que se pode aproveitar outros aspectos tratados na metodologia apresentada. Outra limitação diz respeito ao repositório de palavras geo-interessantes possuir termos apenas nos idiomas português e inglês. A ampliação do repositório de palavras geo-interessantes para outros idiomas seria essencial para ampliar a indexação de sítios de produtores de dados e arquivos geográficos.

Outro ponto interessante seria a definição e implementação de um sistema de perfil para traçar os interesses dos usuários e recomendar novos arquivos geo-interessantes assim que fossem indexados. Para tanto, seria necessário manter um perfil pessoal para cada usuário, que pudesse ser atualizado manualmente ou automaticamente com aprendizado de máquina baseado em padrões de navegação ou respostas para recomendações. Esse sistema já é utilizado com sucesso no Citeseer (Giles *et al.*, 1998).

Nossa ambição é que a utilização do *GeoDiscover* possibilite aos usuários de geoinformação encontrar arquivos relevantes para suas atividades; que se torne uma referência em busca especializada; e amplie, efetivamente, as possibilidades de troca de dados entre produtores

de diferentes regiões geográficas a fim de democratizar os dados existentes nos diversos âmbitos do poder público e privado.

REFERÊNCIAS BIBLIOGRÁFICAS

Abiteboul, S.; Preda, M.; Cobena, G., 2003, **Adaptive on-line page importance computation**, Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, ACM Press.

ADL, 1999, **Alexandria Digital Library Gazetteer**, Map and Imagery Lab, University of California.

Anderson, D. P. BOINC: a system for public-resource computing and storage. In: Fifth IEEE/ACM International Workshop on Grid Computing, 2004 8: p. 4-10.

Anderson, D. P.; Cobb, J.; Korpela, E.; Lebofsky, M.; Werthimer, D. SETI@home: an experiment in public-resource computing. **Communications of the ACM**, v. 45, n. 11, p. 56-61

Arasu, A.; Cho, J.; Garcia-Molina, H.; Paepcke, A.; Raghavan, S. Searching the web. **ACM Transactions on Internet Technology**, v. 1, n. 1, p. 2-43, August 2001.

Baeza-Yates, R. Challenges in the interaction of information retrieval and natural language processing. In: Proceedings of 5th international conference on Computational Linguistics and Intelligent Text Processing (CICLing), 2004, Seoul, Korea. 2945: Springer Berlin / Heidelberg February 2004. p. 445 - 456.

Baeza-Yates, R.; Castillo, C., 2002, **Balancing volume, quality and freshness in web crawling**, Soft Computing Systems – Design, Management and Applications, Santiago, Chile, IOS Press Amsterdam, p. 565-572.

Baeza-Yates, R.; Castillo, C.; Marin, M.; Rodriguez, A., 2005, **Crawling a country: better strategies than breadth-first for web page ordering**, Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, Japan.

Bechhofer, S.; Broekstra, J.; Decker, S.; Erdmann, M.; Fensel, D.; Goble, C.; Harmelen, F. v.; Horrocks, I.; Klein, M.; McGuinness, D.; Motta, E.; Patel-Schneider, P.; Staab, S.; Studer, R. **An informal description of Standard OIL and Instance OIL**. 2000.

Disponível em: http://www.ontoknowledge.org/oil/downl/oil-whitepaper.pdf. Acesso em: 12/07/2006.

Berners-Lee, T.; Hendler, J.; Lassila, O., 2001, **The Semantic Web**, Scientific American.

Boldi, P.; Codenotti, B.; Santini, M.; Vigna, S. UbiCrawler: a scalable fully distributed Web crawler. **Software: Practice and Experience**, v. 34, n. 8, p. 711-726

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

Buckland, M.; Gey, F. The relationship between Recall and Precision. **Journal of the American Society for Information Science**, v. 45, n. 1, p. 12-19 10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L.

Câmara, G.; Casanova, M.; Hemerly, A.; Magalhães, G.; Bauzer-Medeiros, C. **Anatomia de Sistemas de Informação Geográfica**. Curitiba: Sagres, 1996a. 193 p. Disponível em: http://www.dpi.inpe.br/gilberto.

Câmara, G.; Souza, R.; Freitas, U.; Garrido, J. SPRING: Integrating Remote Sensing and GIS with Object-Oriented Data Modelling. **Computers and Graphics**, v. 20, n. 3, p. 395-403, May-Jun 1996. Disponível em: www.dpi.inpe.br/gilberto.

Casanova, M. A.; Câmara, G.; Jr., C. A. D.; Vinhas, L.; Queiroz, G. R. d. **Bancos de dados geográficos**. Curitiba: MundoGEO, 2005. 506 p.

Castillo, C. **Effective web crawling**. 179p p.Department of Computer Science) – Universidad de Chile, Santiago de Chile, 2004.

Cheong, F.-C. **Internet agents: spiders, wanderers, brokers, and bots.** Indianapolis, IN: New Riders Publishing, 1996. 413 p.

Cho, J.; Garcia-Molina, H. Synchronizing a database to improve freshness **ACM SIGMOD Record**, v. 29, n. 2, p. 117-128, June 2000.

___, 2002, **Parallel crawlers**, Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii, USA, ACM Press.

___. Effective page refresh policies for Web crawlers **ACM Transactions on Database Systems**, v. 28, n. 4, p. 390-426

Cho, J.; Garcia-Molina, H.; Page, L. Efficient Crawling Through URL Ordering. In: Seventh International Web Conference (WWW 98). 1998, Brisbane, Australia. April 14-18, 1998.

Cleverdon, C. W. Progress in documentation, evaluation tests of information retrieval systems. **Journal of Documentation**, v. 26, n. 1, p. 55-67

Croft, W. B. Knowledge-based and statistical approaches to text retrieval. **IEEE Intelligent Systems and Their Applications**, v. 8, n. 2, p. 8-12, April 1993.

Densham, I.; Reid, J. A geo-coding service encompassing a geo-parsing tool and integrated digital gazetter service. In: Workshop on the analysis of geographic references, 2003

Dill, S.; Kumar, R.; Mccurley, K. S.; Rajagopalan, S.; Sivakumar, D.; Tomkins, A. Self-similarity in the web **ACM Transactions on Internet Technology** v. 2, n. 3, p. 205-223, August 2002.

Edwards, J.; McCurley, K.; Tomlin, J., 2001, **An adaptive model for optimizing performance of an incremental web crawler**, Proceedings of the 10th international conference on World Wide Web, Hong Kong, Hong Kong, ACM Press.

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

Egenhofer, M.; Frank, A. Prospective views of GIS technologies and applications. In: Simpósio Brasileiro de Geoprocessamento, I, 1990, São Paulo. EPUSP, 1990. p. 95-102.

ESRI. **ESRI Shapefile Technical Description**. Redlands, CA: 1998.

Fonseca, F.; Sheth, A. **The geospatial semantic web**. University Consortium for Geographic Information Science, 2003.

Fonseca, F. T.; Engenhofer, M. J. Knowledge sharing in geographic information systems. In: 1999 Workshop on Knowledge and Data Engineering Exchange, 1999 p. 85.

Foster, I.; Kesselman, C., 1998, **The Globus project: a status report**, Seventh Heterogeneous Computing Workshop, 1998. (HCW 98), Orlando, FL, USA.

Frew, J.; Freeston, M.; Freitas, N.; Hill, L.; Janée, G.; Lovette, K.; Nideffer, R.; Smith, T.; Zheng, Q. The Alexandria Digital Library architecture. **International Journal on Digital Libraries**, v. 2, n. 4, p. 259 - 268, May 2000.

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

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

Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. **Int. Journal of Human-Computer Studies**, v. 43, p. 907-928

Guarino, N.; Giaretta, P. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing 1995, 1995, Amsterdam. IOS Press, p. 25-32.

Heydon, A.; Najork, M. Mercator: A scalable, extensible Web crawler. **World Wide Web**, v. 2, n. 4, p. 219-229, December 1999.

Hirai, J.; Raghavan, S.; Garcia-Molina, H.; Paepcke, A., 2000, **WebBase: a repository of Web pages**, 9th Internationall World Wide Web Conference (WWW9), Amsterdam.

Jones, C. B.; Adbelmoty, A. I.; Finch, D.; Fu, G.; Vaid, S. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. **Lecture Notes in Computer Science**, v. 3234, p. 125-139, Jan 2004.

Jones, K. S.; Willett, P. **Readings in Information Retrieval** Morgan Kaufmann, 1997. 587 p.

Kleinberg, J.; Lawrence, S. The Structure of the Web. **Science**, v. 294, p. 1849-1850, 30 November 2001. Disponível em: www.sciencemag.org.

Kobayashi, M.; Takeda, K. Information retrieval on the web. **ACM Computing Surveys**, v. 32, n. 2, p. 144-173, June 2000.

Koster, M. **Guidelines for Robot Writers**. 1993. Disponível em: http://www.robotstxt.org/wc/guidelines.html. Acesso em: 22/04/2006.

___. Robots in the Web: threat or treat? **Open information interchange spectrum**, v. 2, n. 9, p. 8-18, April 1995. Disponível em: http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html.

___. **A standard for robots exclusion**. 1996. Disponível em: http://www.robotstxt.org/wc/exclusion.html. Acesso em: 25/04/2006.

Laender, A. H. F.; Borges, K. A. V.; Carvalho, J. C. P.; Medeiros, C. B.; Silva, A. S. d.; Davis, C. A. Integrating web data and geographic knowledge into spatial databases. In: Manolopoulos, Y.; Papadopoulos, A.; Vassilakopoulos, M. (Ed.). **Spatial Databases: Technologies, Techniques and Trends**. Hershey, Pennsylvania, USA, 2004.

Lawrence, S.; Giles, C. L. Accessibility of information on the web. **Nature**, v. 400, n. 6740, p. 107-107.

Lee, D. L. **Performance evaluation of information retrieval systems**. Lawrence: University of Kansas Powerpoint. Disponível em: http://www.cs.ust.hk/~dlee/336/.

Lessing, L. **Code and Other Laws of Cyberspace** Basic Books, 1999. 297 p.

Li, Y. Toward a qualitative search engine. **IEEE Internet Computing**, v. 2, n. 4, p. 24-29, Jul/Aug 1998.

Masolo, C.; Gangemi, A.; Guarino, N.; Oltramari, A.; Schneider, L. **The WonderWeb Library of Foundational Ontologies**. Padova: LADSEB-Cnr, 2002. 36 p. (2001-33052).

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

Miller, R. C.; Bharat, K. SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers. In: Proceedings of the Seventh International World Wide Web Conference (WWW7), 1998, Brisbane, Australia. 30: ISDN Systems, April 1998. p. 119-130.

Moore, D.; Periakaruppan, R.; Donohoe, J. Where in the world is netgeo.caida.org? In: INET-2000 The 10th Annual Internet Society Conference, 2000

Najork, M.; Wiener, J. L., 2001, **Breadth-first crawling yields high-quality pages**, Proceedings of the 10th international conference on World Wide Web Hong Kong, Hong Kong, ACM Press.

OGC. **The OpenGIS specification model**. 1999. Disponível em: www.opengeospatial.org/specs/.

Page, L.; Brin, S.; Motwani, R.; Winograd, T. **The PageRank Citation Ranking: Bringing Order to the Web**. 1999. Disponível em: <http://dbpubs.stanford.edu/pub/1999-66>.

Pinkerton, B., 1994, **Finding What People Want: Experiences with the WebCrawler**, Second International WWW Conference.

Porter, M. The Porter Stemming Algorithm. Cambridge: 2006.

Raghavan, V. V.; Jung, G. S.; Bollmann, P. A critical investigation of Recall and Precision as measures of retrieval system performance **ACM Transactions on Information Systems**, v. 7, n. 3, p. 205-229, July 1989.

Risvik, K. M.; Michelsen, R. Search Engines and Web Dynamics. **Computer Networks**, v. 39, p. 289-302, June 2002.

Ritter, N.; Ruth, M. The GeoTiff data interchange standard for raster geographic images. **International Journal of Remote Sensing** v. 18, n. 7, p. 1637-1647, May 1997.

Silberschatz, A.; Korth, H. F.; Sudarshan, S. **Database system concepts**. 4th edition McGraw-Hill, 2002. 1129 p.

Silva, A. S.; Veloso, E. A.; Golgher, P. B.; Ribeiro-Neto, B.; Laender, A. H. F.; Ziviani, N. Cobweb – a crawler for the Brazilian web. In: String Processing and Information Retrieval (SPIRE), 1999, Cancun, Mexico. IEEE CS Press, p. 184-191.

Silva, M. J.; Martins, B.; Chaves, M.; Cardoso, N.; Afonso, A. P., 2004, **Adding Geographic Scopes to Web Resources**, ACM SIGIR 2004 Workshop on Geographic Information Retrieval, Sheffield, UK.

Taufer, M.; An, C.; Kerstens, A.; III, C. L. B. Predictor@Home: A "Protein Structure Prediction Supercomputer" Based on Public-Resource Computing. In: 19th IEEE International Parallel and Distributed Processing Symposium, 2005 4-8 April 2005. p. 8pp.

Wiederhold, G. Interoperation, Mediation and Ontologies. In: International Symposium on Fifth Generation Computer Systems (FGCS94), 1994, Tokyo, Japan. W3: ICOT, p. 33-48.