UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICAL
COORDENAÇÃO DE PÓS-GRADUAÇÃO EM INFORMÁTICA

GeoSEn: um Motor de Busca com Enfoque Geográfico

Cláudio Elízio Calazans Campelo

Campina Grande, Paraíba, Brasil
Outubro de 2008

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
COORDENAÇÃO DE PÓS-GRADUAÇÃO EM INFORMÁTICA

GeoSEn: um Motor de Busca com Enfoque Geográfico

Cláudio Elízio Calazans Campelo

Dissertação submetida à coordenação do curso de pós-graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I, como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Cláudio de Souza Baptista
(Orientador)

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Sistemas de Informação e Banco de Dados

Campina Grande, Paraíba, Brasil
Outubro de 2008

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

C193g
2008 Campelo, Cláudio Elízio Calazans.
GeoSEn: um motor de busca com enfoque geográfico / Cláudio Elízio Calazans Campelo. ─ Campina Grande, 2008.
148 f. : il.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática.
Referências.
Orientador: Dr. Cláudio de Souza Baptista.

1. Recuperação de Informação. 2. Sistemas de Informações Geográficas. 3. Motores de Busca para a Web. 4. Banco de Dados Espaciais. I. Título.

CDU -004.738.52 (043)

Resumo

A Recuperação de Informação Geográfica (do inglês, GIR - *Geographic Information Retrieval*) é uma área de pesquisa recente e que vem se mostrando bastante atraente. Diante do rápido crescimento do volume de informações disponibilizadas na Web, conseguir recuperar, de maneira simples e eficiente, documentos que atendam às necessidades cada vez mais específicas dos usuários é o grande desafio no campo da Recuperação de Informação (RI) hoje. Os motores de busca geográficos para Web são especializações dos motores de busca tradicionais e visam adicionar a estes a capacidade de identificar o contexto geográfico dos elementos da Web (e.g., textos, imagens, vídeos) e indexá-los segundo tais características. As pesquisas no campo da GIR dividem-se em sub-áreas que têm como principais objetivos: o desenvolvimento de métodos para detecção de referências geográficas nos elementos da Web; a elaboração de modelos de representação do escopo geográfico destes elementos, ou seja, do conjunto de lugares aos quais estes estão associados; a construção de estruturas de dados e algoritmos eficientes para indexação espaço-textual; a concepção de modelos para geração do *ranking* de relevância que combinem as dimensões textual e espacial; a criação de interfaces com o usuário e mecanismos de busca que permitam representar as consultas segundo tais dimensões e recuperar os elementos mais relevantes. Este trabalho apresenta um modelo elaborado para construção de um sistema de GIR, com foco nos processos de modelagem do escopo e de geração do *ranking* de relevância geográficos. Para validação destes, foi desenvolvido um protótipo de um motor de busca para Web, chamado de GeoSEn - *GEOgraphic Search ENgine*, que integra os principais mecanismos previstos em um motor de busca geográfico para Web.

Abstract

Geographic Information Retrieval (GIR) is a recent research area which has become very attractive. Regarding the fast growth of the amount of information available on the Web, the provision of simple and efficient retrieval tools to be used by different user needs is currently one of the most challenge in the field of Information Retrieval (IR). Geographic Web search engines are specializations of the standard Web search engines, which add to them the ability to identify the geographic context of Web resources (e.g., texts, images, movies) and to index them using the spatial features. Research on GIR may be categorized in a few subareas which have specific objectives: to develop methods to detect geographic references in Web resources; to elaborate representation models for these resources' geographic scope, that is, the set of places the documents are associated to; to build efficient data structures and algorithms for spatio-textual indexing; to conceive models for relevance ranking, which may combine both spatial and textual dimensions; and to design user interfaces and search mechanisms which allow to represent user queries by these dimensions and then to retrieve the most relevant documents. This work presents an elaborated model for a GIR system building, emphasize both geographic scope modeling and relevance ranking generation processes. In order to validate the proposed model, a geographic Web search engine prototype, called GeoSEn – GEOgraphic Search ENgine, was built. GeoSEn integrates the main functions contained in a geographic Web search engine.

Agradecimentos

À minha esposa Carol, pelo amor, companheirismo, dedicação e pela paciência em me dividir com o mestrado. Fatores essenciais para que alcançasse este objetivo.

À minha mãe querida, pelo incentivo incondicional, pelo amor e por me estimular aos estudos desde os primeiros anos de vida.

Ao meu paizão, pelo seu carinho, apoio, por seus exemplos de determinação e competência, e por sempre ter como prioridade o investimento na educação dos seus filhos, o que foi fundamental para que eu pudesse chegar até aqui.

Ao meu avô Felinto, pelo exemplo de como um homem deve ser. Às minhas avós, pela imensa afeição.

Às minhas irmãs queridas, pelo amor e amizade.

Ao professor Cláudio Baptista, pela amizade e pelo excelente trabalho de orientação.

Aos amigos Daniel, Hugo, Yuri, Welmisson, Fábio e tantos outros que contribuíram para que os dias em Campina Grande fossem mais divertidos, com bate-papos, futebol e cervejadas. Ainda ao Hugo pelos “galhos quebrados” quando estive longe de Campina.

Aos demais familiares e amigos, que são fundamentais para a minha felicidade.

À Luciana e Ricardo, pela colaboração neste trabalho. A este também por manter o GeoSEn vivo através de sua pesquisa de mestrado.

A todos os companheiros de trabalho do LSI, pelas experiências compartilhadas.

A todos os professores do DSC, pelos conhecimentos transmitidos desde o curso de graduação.

Aos funcionários da COPIN, em especial à Aninha, pela sua atenção, simpatia e prestatividade.

A CAPES, pelo apoio financeiro no período em que este foi necessário.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Códigos

Lista de Equações

Lista de Siglas

API	- Application Programming Interface (Interface de Programação de Aplicativos).
GIR	- Geographic Information Retrieval (Recuperação de Informação Geográfica).
GPS	- Global Positioning System (Sistema de Posicionamento Global).
HTML	- HyperText Markup Language (Linguagem de Marcação de Hipertexto).
HTTP	- Hypertext Transfer Protocol (Protocolo de Transferência de Hipertexto)
JPEG	- Joint Photographic Experts Group (Grupo de Especialistas em Fotografia).
OWL	- Web Ontology Language (Linguagem Ontológica para Web).
PDF	- Portable Document Format (Formato de Documento Portável).
RI	- Recuperação de Informação.
RSS	- Really Simple Syndication (Distribuição Realmente Simples).
SGBD	– Sistema de Gerenciamento de Bancos de Dados.
URL	- Uniform Resource Locator (Localizador de Recursos Universal).
XML	- Extensible Markup Language (Linguagem de Marcação Extensível).

Capítulo 1
Introdução

Diante da vasta quantidade de informações disponíveis na Web, conseguir recuperar, de maneira simples, informações que atendam precisamente a determinadas necessidades é uma tarefa cada vez mais necessária e requisitada entre os usuários. Grandes contribuições foram obtidas na área de Recuperação da Informação (RI) desde os anos 60. No entanto, com o grande crescimento da Web, as pesquisas na área se intensificaram, endereçando novos problemas cujas soluções vêm sendo possibilitadas pelo avanço, paralelamente, de outras áreas da ciência da computação.

De maneira bastante superficial, o funcionamento dos tradicionais sistemas de busca disponíveis hoje na Web consiste em recuperar documentos que contenham algumas palavras-chave especificadas pelo usuário. No entanto, dentre as diversas possibilidades de busca, é de se esperar que o usuário necessite de um processo que envolva análises mais sofisticadas, como por exemplo, com contextualização temporal ou espacial. Quando se trata de contextualização temporal, por exemplo, um sistema pode considerar que, dentre alguns documentos similares, os documentos mais recentes sejam mais relevantes.

O foco deste trabalho é a recuperação de informação considerando o âmbito textual e geográfico. Visto que grande parte dos documentos disponíveis na Web possui algum tipo de contextualização geográfica, torna-se possível, através de um tratamento especial a este tipo de informação, satisfazer a determinados requisitos que, em sistemas tradicionais de busca na Web, são muito difíceis ou mesmo impossíveis de serem atendidos.

Para muitas buscas efetuadas pelos usuários, os resultados considerados mais interessantes são aqueles que estão relacionados à região onde estes estão localizados ou a regiões próximas. Por exemplo, em uma pesquisa por emprego, em geral, as informações mais

relevantes são as encontradas na cidade onde o interessado reside ou, como alternativas, nas cidades vizinhas. Outra aplicação interessante é o turismo, onde o usuário pode procurar por hotéis e locadoras de veículos especificando o local de destino.

Nos sistemas de busca baseados em palavras-chave, por exemplo, uma página contendo a frase "...Com a chegada da empresa em Campina Grande, mil novas vagas serão abertas para programadores java..." não seria recuperada através da pesquisa "vagas programadores paraíba", salvo se em algum outro trecho do texto contido na página fosse encontrada também a palavra "paraíba". Isto acontece porque a expressão "campina grande" é tratada como qualquer outra, e não como uma localização geográfica. Em um sistema de busca geográfico seria esperado que a página do exemplo fosse retornada, visto que este teria a informação que a expressão "campina grande" se refere a uma cidade e que esta está localizada em um estado que pode ser referenciado pelo termo "paraíba". No sistema tradicional, este resultado só seria possível se o usuário repetisse a consulta para todas as possíveis sub-regiões (e.g. municípios) do estado da Paraíba.

1.1. Objetivos

1.1.1. Objetivos Gerais

O objetivo deste trabalho é o desenvolvimento de um protótipo de um sistema de recuperação de informação geográfica para Web, com enfoque no mecanismo de detecção do escopo geográfico contido nos documentos Web, bem como na sua modelagem e representação, permitindo que estes sejam recuperados com base em informações textuais de espaciais especificadas pelos usuários. O sistema proposto é chamado de GeoSEn, uma sigla para GEOgraphic Search ENgine.

1.1.2. Objetivos Específicos

OE1 - Detectar referências a localizações geográficas nos documentos. Após uma determinada página ser capturada pelo Web coletor, esta é analisada gramaticalmente (do inglês, "parsed") em busca de termos que possam ser mapeados a algum nome de lugar, e.g., nomes de cidades, códigos postais, telefones, gentílicos, etc. Estes termos são chamados de termos candidatos, e os resultados destes mapeamentos, de referências geográficas. Em seguida, esses nomes de lugares são associados a coordenadas geográficas, utilizadas para manipulação de informações espaciais internamente pelo sistema. Em uma mesma página podem ser encontradas referências a uma ou mais localidades. Este mecanismo deve ser capaz de resolver diferentes tipos de ambiguidades, como por exemplo, a existência de lugares distintos com o mesmo nome. Algumas heurísticas conhecidas para o processo de detecção dos termos devem ser aperfeiçoadas e utilizadas em conjunto com as novas heurísticas propostas. Além de verificar o conteúdo dos documentos em busca de referências geográficas, o sistema deve realizar análise das URLs relacionadas a estes documentos.

OE2 - Modelar o escopo geográfico dos elementos da web. De posse das referências geográficas previamente detectadas, o sistema deve ser capaz de fazer inferências sobre o escopo geográfico do documento, ou seja, sobre as regiões às quais o documento se refere. Para inferir algo sobre o escopo geográfico dos documentos com base nas referências geográficas encontradas, utilizam-se heurísticas e algoritmos especializados, combinando, além de várias propriedades dos termos encontrados, diferentes relacionamentos espaciais entre as regiões analisadas. Um dos fatores a ser considerado por tais heurísticas é o padrão de distribuição (ou de espalhamento) das referências.

OE3 – Realizar indexação espaço-textual. Após os documentos serem capturados e analisados segundo suas características geográficas, estes devem ser organizados em uma base única utilizando técnicas de indexação espaço-textual, permitindo que sejam recuperados de forma eficiente a partir de dados espaciais e textuais provenientes do módulo de busca.

OE4 – Recuperar os documentos utilizando diferentes operações espaciais. Uma vez indexados os documentos, deve ser possível recuperá-los especificando-se algumas informações textuais e espaciais que caracterizem os documentos desejados. O sistema deve

disponibilizar ao usuário diferentes operadores espaciais, como por exemplo, os operadores de continência, adjacência e distância, bem como suas respectivas negações.

OE5 – Elaborar um ranking de relevância espaço-textual. Os documentos recuperados pelo módulo de busca devem ser ordenados segundo um ranking de relevância que avalie e pondere a importância do documento em relação aos argumentos textual e espacial da busca submetida ao sistema. Deve ser possível ao administrador do sistema configurar o peso de cada uma destas perspectivas no cálculo do valor final de relevância.

OE6 - Possuir interface multi-modo. A interação dos usuários com o sistema se dá através de uma interface multi-modo, contendo um módulo para entrada e visualização de dados textuais e um mapa interativo para entrada de dados geográficos. Em uma interface como esta deve ser possível selecionar visualmente as localidades a serem processadas pela consulta, cujas podem ser pré-definidas (e.g., uma cidade) ou mesmo uma seleção livre (e.g., seleção retangular). Nesta interface, além dos dados textuais e geográficos da consulta, o usuário especifica também o operador espacial a ser utilizado e seus respectivos parâmetros, caso possuam (e.g., um valor em quilômetros para o operador de distância). Disponibiliza-se ainda uma tela para eliminação de ambiguidades, onde é possível selecionar a localidade exata de interesse, no caso de ter sido especificada textualmente uma localidade ambígua.

1.2. Relevância

Alguns trabalhos conhecidos na literatura tratam da atribuição de escopo geográfico às páginas da Web e, em geral, os métodos sugeridos são baseados em heurísticas. Entretanto, a maioria dos algoritmos e heurísticas propostos possui algumas limitações e diversas possibilidades de aprimoramento. Neste trabalho, propõe-se: o aperfeiçoamento de algumas heurísticas já conhecidas, bem como um conjunto de novas heurísticas para o processo de detecção de referências geográficas em documentos da Web; e um modelo para a composição e representação do escopo geográfico dos documentos e para o cálculo dos valores de relevância geográfica das localidades associadas a cada documento.

1.3. Organização Estrutural

O restante desta dissertação está organizado da seguinte forma: o Capítulo 2 discorre sobre os fundamentos da recuperação de informação e dos motores de busca para Web, bem como sobre os fundamentos e estado da arte no campo da recuperação de informação geográfica. O Capítulo 3 apresenta as principais contribuições na área e as pesquisas mais importantes que exploram o tema. No Capítulo 4, descrevem-se os métodos e técnicas elaborados neste trabalho; apresenta-se ainda o protótipo desenvolvido para validação das idéias propostas. No Capítulo 5, são mostrados os experimentos realizados para avaliação do protótipo desenvolvido. Por fim, discutem-se no Capítulo 6 as limitações observadas e as possibilidades de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

O volume de informações disponíveis na Web vem crescendo de forma muito rápida e, diante disso, diversas preocupações e desafios de como tratar, organizar e recuperar estas informações têm sido lançadas, caracterizando os objetivos principais de um sistema de busca para Web (*Web search engine*). A Web adiciona novas características e dificuldades ao processo de recuperação da informação (RI), seja pela forte heterogeneidade entre os documentos disponíveis, seja pela forma como esses documentos estão disponibilizados. Este novo contexto proporcionou a intensificação e o avanço das pesquisas em RI, onde novas soluções se fizeram possíveis com o avanço de outras áreas da computação, como a de sistemas distribuídos e computação paralela.

Grande parte das informações dispersas na Internet possui algum tipo de contexto geográfico. Podem fazer parte deste contexto, por exemplo, o local onde a informação foi criada, os locais aos quais os conteúdos se referem, onde estão os maiores interessados por estas informações, etc. No entanto, os sistemas tradicionais de recuperação de informação não consideram este contexto nos processos de organização e recuperação das informações. Em geral, nos sistemas de busca disponíveis na Web, o usuário fornece algumas palavras-chave e o sistema retorna os documentos que as contêm. Caso alguma destas seja o nome de um lugar, por exemplo, o sistema não tratará de forma especial esta requisição, ou seja, recuperará os documentos contendo esta palavra, como outra qualquer.

A adição de contextualização geográfica nos processos envolvidos com a recuperação de informação pode proporcionar resultados mais interessantes ao usuário, ou mesmo fazer com que uma simples consulta substitua múltiplas consultas a um sistema de busca tradicional para se obter um resultado similar. Ainda são poucas as pesquisas relacionadas ao tema, mas

já surgem alguns sistemas resultantes de iniciativas acadêmicas [1][2] e comerciais [3][4][5][6].

O restante deste capítulo está estruturado da seguinte forma: inicialmente, apresentam-se os principais conceitos de RI, alguns de seus modelos clássicos ainda utilizados como base para muitos sistemas modernos e os principais componentes de um sistema de busca para Web. Em seguida, descrevem-se as características de um sistema de recuperação de informação geográfica, discutindo-se sobre as atividades necessárias e as dificuldades envolvidas na integração destas novas soluções às arquiteturas e interfaces de sistemas de busca tradicionais.

2.1. Fundamentos de RI

Há de se distinguir os sistemas de recuperação de informações dos sistemas de recuperação de dados. Conforme Rijsbergen [12], os sistemas de recuperação de dados visam recuperar todos os objetos que satisfazem precisamente a condições claramente definidas, expressas através de linguagens baseadas, por exemplo, em expressões regulares ou em álgebra relacional. Nestes sistemas, o resultado é fruto de uma busca completa e exaustiva. A recuperação de informações traz dificuldades intrínsecas ao conceito de "informação", como a dificuldade de determinar a verdadeira necessidade do usuário e atender às suas expectativas através de um subconjunto relevante de documentos que fazem parte do acervo do sistema. Em um sistema de RI, faz-se necessária uma interpretação sintática e semântica do conteúdo dos documentos e os resultados das consultas podem ser imprecisos. Baeza-Yates et al [7] exibem um modelo de recuperação de informação definido como uma quádrupla [*D, Q, F, R*(q_i, d_j)], onde:

1. **D** é o conjunto de visões lógicas dos documentos do sistema;
2. **Q** é o conjunto de visões lógicas das necessidades dos usuários (*queries*);
3. **F** é um *framework* para modelar **D**, **Q** e seus relacionamentos;
4. **R($\mathbf{q_i}$, $\mathbf{d_j}$)** é uma função de *Ranking* que associa um número real a uma *query* $\mathbf{q_i} \in \mathbf{Q}$ e uma visão lógica de um documento $\mathbf{d_j} \in \mathbf{D}$.

A seguir, apresentam-se alguns conceitos importantes relacionados aos sistemas de Recuperação de Informação.

2.1.1. Relevância

Segundo Kowalsky et al [8], em geral, da perspectiva do usuário, "relevância" e "necessidade" são sinônimos. Ou seja, é esperado que um sistema de RI retorne, dentre os resultados para uma determinada consulta do usuário, o maior número possível de itens relevantes (que realmente sejam de interesse do usuário).

Diversas técnicas foram desenvolvidas com o objetivo de aumentar o grau de relevância dos resultados de uma consulta. Em geral, procura-se constituir um índice de similaridade entre o conjunto de identificadores dos documentos e o conjunto de termos da consulta. Com base nesta medida de similaridade, um *ranking* de documentos pode ser recuperado e apresentado de acordo com uma consulta solicitada. Nos motores de busca para Web, além deste grau de similaridade, adotam-se outras técnicas para construção do *ranking* de relevância, que consideram, por exemplo, a estrutura de ligações (*links*) entre as páginas da Web, a exemplo do modelo descrito em [9][10].

2.1.2. Avaliação de RI

Duas métricas são geralmente associadas a um sistema de recuperação de informação, com o objetivo de mensurar a eficiência do processo de avaliação da relevância de documentos: precisão (*precision)* e revocação (*recall)*. Conforme mencionado, um documento retornado por um sistema destes pode ser relevante ou não-relevante para o usuário. Além disso, neste processo de pesquisa, há duas possibilidades em relação à recuperação de um documento: pode ter sido *recuperado* ou *não-recuperado*, através de uma busca específica. A Figura 2.1 apresenta os quatro possíveis efeitos de uma consulta sobre o espaço total de documentos.

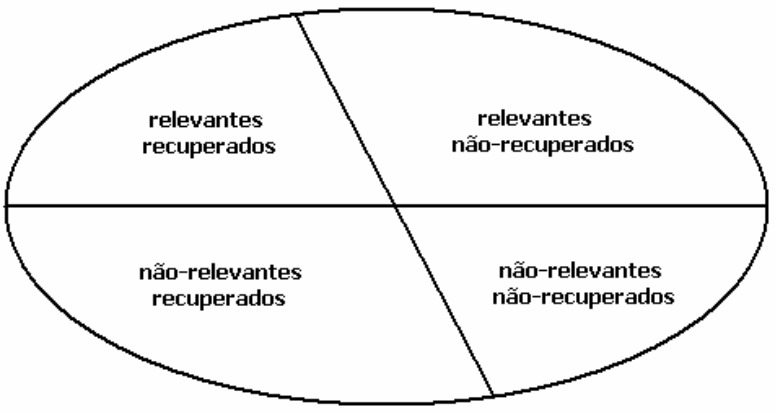


Figura 2.1 – Efeito de uma pesquisa no espaço total de documentos [8].

Com isso, as duas medidas citadas anteriormente podem ser expressas segundo a Equação 2.1:

$$Precisão = \frac{Número\ relevantes\ recuperados}{Total\ recuperados} \qquad \text{Equação 2.1}$$

$$Revocação = \frac{Número\ relevantes\ recuperados}{Total\ relevantes\ existentes}$$

Precisão mede a eficiência da recuperação de informação para uma consulta específica, considerando apenas os resultados obtidos. Por exemplo, se a precisão é de 70%, significa que houve 30% de itens não-relevantes recuperados. Já a revocação serve para demonstrar, em uma determinada consulta, quão hábil foi a recuperação de itens considerados relevantes, considerando todos os documentos possíveis de serem recuperados. Por exemplo, uma revocação de 20% significa que houve 80% de itens relevantes que não foram recuperados.

Apesar da popularidade das duas medidas apresentadas, elas nem sempre são as mais apropriadas para medir a performance de uma RI. Citam-se algumas outras medidas alternativas, como por exemplo, a chamada de Média Harmônica (*Harmonic Mean*) [11] e a Medição E (*E Measure*) [12]. Ainda, uma vez que diferentes usuários podem ter diferentes interpretações sobre qual documento é relevante e qual não é, surgiram as medidas orientadas ao usuário, como por exemplo, as propostas por Korfhage [13]: razão de cobertura (*coverage*

ratio), razão de novidades (*novelty ratio*), revocação relativa (*relative recall*) e esforço da revocação (*recall effort*).

2.1.3. Operações Textuais

Conforme Baeza-Yates et al [7], os computadores modernos vêm possibilitando a representação de documentos através de seu conjunto completo de palavras. Neste caso, afirma-se que tal sistema de RI adota o método de representação (ou visão lógica) de documentos chamado de *full text*. De acordo com os autores, esta forma de representação sem dúvida é a forma mais completa de representação. No entanto, para grandes coleções de documentos, pode ser necessário diminuir esse conjunto de palavras representativas, com o objetivo de reduzir os custos computacionais. Esse processo de redução, em geral, dar-se através da eliminação de *stopwords* (como artigos e preposições), do uso de *stemming* (que reduz palavras distintas a suas raízes gramaticais comuns) e da identificação de grupos de substantivos (que elimina adjetivos, verbos e advérbios). Em seguida, processos de compressão podem ser empregados. Estas operações são chamadas de operações textuais (ou transformações). Com estes procedimentos, reduz-se a visão lógica do documento a um conjunto de termos indexáveis. De forma análoga, para as consultas dos usuários (em geral especificadas através de um conjunto de palavras), uma visão lógica pode ser concebida utilizando-se tais operações.

2.1.4. Indexação

Índices são largamente utilizados em sistemas de RI, fundamentais para promover agilidade no processo de busca quando se manipula uma grande coleção de documentos. Baeza-Yates et al [7] citam as três principais técnicas de indexação: arquivos invertidos (*inverted files*), arrays de sufixo (*suffix arrays*) e arquivos de assinatura (*signature files*). Dentre estes, destaca-se o índice em arquivo invertido, atualmente mais difundido e utilizado nas aplicações de RI.

Araújo et al [14] definem um arquivo invertido como um mecanismo orientado a palavras para indexar coleções textuais e promover maior desempenho na atividade de busca. Os autores descrevem o índice contendo dois elementos: vocabulário e ocorrências. O vocabulário é o conjunto de diferentes palavras no texto. Para cada palavra é armazenada uma lista contendo todas as posições em que a palavra aparece no texto. O conjunto de todas estas listas é chamado de ocorrências. A Figura 2.2 mostra um exemplo de um índice em arquivo invertido formado para um pequeno texto.

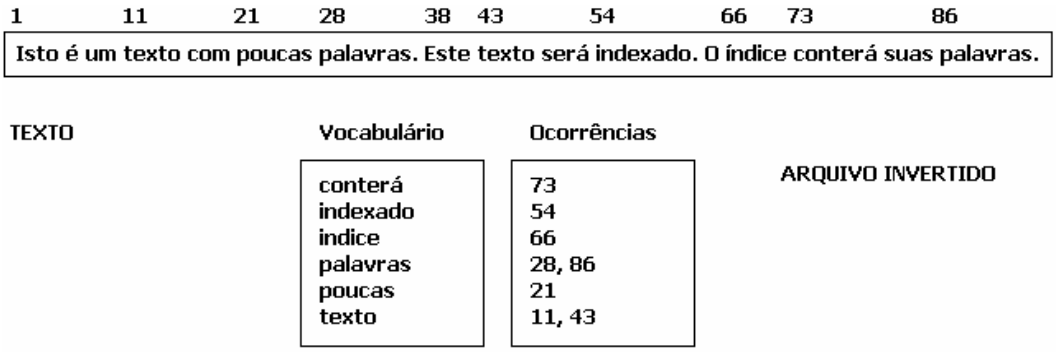


Figura 2.2 – Arquivo invertido formado a partir de um pequeno texto.

2.1.5. Arquitetura

A Figura 2.3 apresenta uma arquitetura básica de um software genérico de RI. Nesta simples arquitetura, exibida por Baeza-Yates et al [7], primeiramente define-se o banco de dados: os documentos a serem usados, as operações textuais a serem utilizadas, etc. Em seguida, definida a visão lógica dos documentos, constrói-se o índice (geralmente um arquivo invertido). Construído o índice, o processo de busca pode ser iniciado. Os autores propõem o seguinte fluxo de funcionamento:

- O usuário especifica sua pesquisa, que é analisada por um *parser* e transformada utilizando as mesmas operações textuais utilizadas nos documentos;
- A consulta gerada (representação do sistema para a pesquisa do usuário) é processada para obter os documentos que a satisfazem;

- Os documentos selecionados são ordenados segundo critérios de relevância e retornados ao usuário na forma de *ranking*;
- O usuário examina os resultados e, neste ponto, pode iniciar um ciclo de *feedback*, onde o sistema utiliza os documentos recuperados para alterar a consulta;

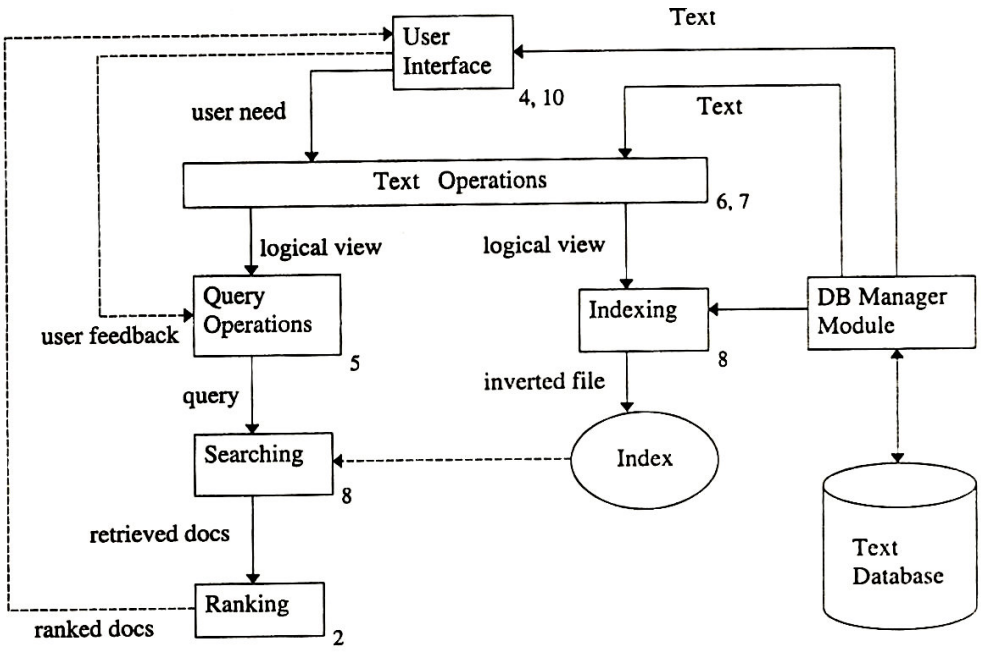


Figura 2.3 - Arquitetura básica de um sistema de RI [7].

2.1.6. Modelos Clássicos

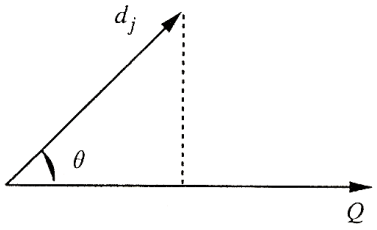
Os modelos clássicos de RI consideram que um documento é descrito por um conjunto de termos indexados. Alguns desses modelos consideram que um termo pode ter mais representatividade do que outro do mesmo documento. Essa representatividade é chamada de *peso* (*w*) e pode ser descrita por um valor numérico. Em cada modelo, estes pesos podem ser ou não considerados como mutuamente independentes. Os modelos clássicos mais conhecidos são o booleano [14], o probabilístico [16] e o vetorial [17][18][19].

O booleano, discutido por Wartick [14], é um simples modelo baseado na teoria dos conjuntos e na álgebra booleana. Neste modelo, uma *query* é uma expressão *booleana*

convencional composta de termos ligados por conectivos *NOT*, *AND*, ou *OR*. Já os pesos relacionados aos termos são binários, ou seja, $w \in \{0,1\}$. A principal vantagem deste modelo é seu formalismo claro e sua simplicidade. Já sua desvantagem é o fato de um documento ser considerado apenas relevante ou não-relevante, não permitindo relevância parcial, ocasionando, por vezes, resultados muito grandes ou muito pequenos.

O modelo probabilístico, introduzido por Robertson e Jones [16], tenta estimar, para uma dada consulta q_i e um documento d_j na coleção, a probabilidade do usuário considerar o documento d_j relevante. Para isso, considera-se que existe um subconjunto dos documentos da coleção contendo apenas documentos relevantes. Este subconjunto é chamado de conjunto ideal. Dada uma descrição inicial deste conjunto ideal, um subconjunto dos documentos retornados como resultado para a consulta é considerado relevante. Esses documentos são então utilizados para refinar a descrição do conjunto ideal. Este processo é realizado repetidamente até obter-se uma aproximação do conjunto desejado. Uma vantagem desse modelo é a possibilidade de formar um *ranking* de acordo com a probabilidade de relevância dos documentos; como desvantagem, aponta-se, dentre outros, o fato de não considerar a frequência em que um termo ocorre em um documento como fator para mensurar a representatividade do termo para o documento.

O modelo vetorial, proposto por Salton [17][18][19], é o modelo de maior aceitação dentre os pesquisadores e o mais utilizado pelas aplicações atuais de RI. Neste modelo, associa-se a cada termo de indexação k_i, em um documento dj, um peso $w_{i,j} \geq 0$, que quantifica a correlação entre os termos e o documento. Também são atribuídos pesos aos termos da consulta (*q*). O modelo vetorial visa quantificar a similaridade entre os documentos e uma consulta realizada pelo usuário, através de uma correlação entre os vetores $\vec{q}=(w_{1,q}, w_{2,q,\ldots}, w_{t,q})$ e $\vec{d}_j=(w_{1,j}, w_{2,j,\ldots}, w_{t,j})$, onde t é o número total de termos indexados no sistema. Os resultados são dados em ordem decrescente de similaridade, considerando também documentos que satisfazem parcialmente a consulta. A correlação entre $\vec{q}$ e $\vec{d}_j$ é dada pela equação abaixo, resultando no cosseno do ângulo entre os vetores, que determina a similaridade entre eles, conforme ilustrado na Figura 2.4.



$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

Figura 2.4 – Cosseno de θ representa sim(d_j, q) [7].

O peso de um termo em um documento pode ser calculado de diversas formas. Estes métodos de cálculo de peso geralmente se baseiam no número de ocorrências do termo no documento (frequência). Uma das formas de se calcular o peso, chamada de abordagem TF-IDF, tenta balancear características em comum nos documentos (*intra-cluster*) e características para fazer a distinção entre os documentos (*inter-cluster*), combinando, respectivamente, as medidas de Frequência do Termo (*Term Frequecy*), ou TF, que mede a freq uência de um termo k_i dentro de um documento d_j e outra medida chamada de Frequência Inversa de Documentos (*Inverse Document Frequency*), ou IDF, que mede a frequência inversa de um termo k_i na coleção de documentos. Esta cálculo é exibido na Equação 2.2, onde N é o número total de documentos no sistema, n_i o número de documentos nos quais o termo k_i aparece, $freq_{i,j}$ o número de vezes em que o termo k_i é mencionado no texto de d_j e o máximo é calculado sobre todos os termos mencionados no texto de d_j.

$$TF_{i,j} = \frac{freq_{i,j}}{max_l \quad freq_{l,j}}$$

$$IDF_i = Log \frac{N}{n_i}$$

Equação 2.2

$w_{i,j} = TF_{i,j} \times IDF_i$

2.2. Motores de Busca para Web

De acordo com Page et al [20], documentos disponíveis na Web são extremamente heterogêneos em diversos aspectos, sejam pela utilização de diversas linguagens (humana ou de programação), pelo emprego de variados idiomas, pelo formato utilizado para representação do conteúdo (texto puro, HTML, PDF, imagem, vídeo) e, ainda, se este foi produzido por ser humano ou por máquina (*logs* de sistemas, por exemplo), além de outros fatores externos ao conteúdo do documento, como por exemplo, a reputação da fonte, a frequência de atualização, dentre outras características. Ainda, tem-se que não há nenhum controle sobre as pessoas que produzem esses documentos e onde o armazenam, além de usuários e companhias que se tornaram especializados em manipular sistemas de busca, na maioria das vezes com interesses comerciais. Esses e outros fatores adicionam um grau de complexidade mais elevado a sistemas de recuperação de informação quando estes são voltados para Web.

Um motor de busca para Web é um ótimo exemplo de aplicação prática de um sistema de recuperação de informação. Atualmente os serviços de busca de informações são, juntamente com os correios eletrônicos, os serviços mais utilizados na Web. Diversos motores de busca estão disponíveis na Web, como por exemplo, o Google (www.google.com), o Yahoo! (www.yahoo.com) e o Altavista (www.altavista.com), sendo todos baseados em algoritmos de RI.

Os sistemas de busca consistem basicamente de três partes:

- O *Web crawler* (também conhecido como *spider*, *walker* ou robô) que localiza e captura as páginas da Web;
- O *indexador*, que indexa as páginas capturadas pelo *robô* de acordo com os termos (palavras) de seu conteúdo e armazena o índice de termos resultante em uma base de dados;
- O *processador de consultas*, que confronta a consulta do usuário (*query*) com o índice e retorna um resultado contendo os documentos que consideram mais relevantes.

2.2.1. Robôs da Web

Segundo Heydon e Najork [23], os robôs da Web são tão antigos quanto a própria Web. São aplicações especializadas que coletam páginas a serem indexadas. O primeiro robô, Matthew Gray's Wanderer, foi desenvolvido em 1993. Vários trabalhos sobre os coletores da Web (*Web crawling*) foram apresentados nas primeiras duas conferências da Web. Nesse tempo a Web não tinha a dimensão que tem hoje, desta forma, tais sistemas não endereçavam os problemas de escalabilidade que existem nos robôs da Web de hoje.

Todas os sistemas de busca populares utilizam robôs que devem alcançar as porções substanciais da Web. Heydon e Najork [23] comentam que, devido à competitividade entre os produtores dos sistemas de busca, os projetos desses robôs não têm sido publicamente descritos. No entanto, existem duas exceções, o *Google Crawler* e o *Internet Archive Crawler*, que possuem alguma descrição disponível na literatura, porém muito concisas. A Figura 2.5 mostra uma arquitetura básica para um robô da Web, propostos pelos autores.

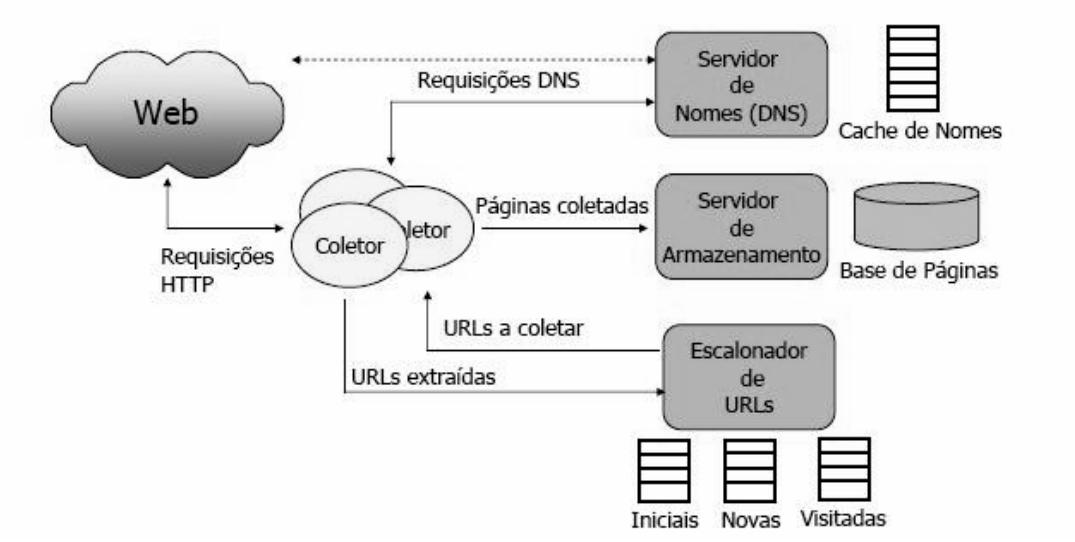


Figura 2.5 – Arquitetura Básica de um robô da Web [12].

Nesta arquitetura, os *coletores* são responsáveis pela requisição de páginas aos servidores HTTP, por extrair os links das páginas recebidas e enviá-las ao escalonador e por requisitar deste uma ou mais URLs a serem coletadas. O *servidor de armazenamento* é o

responsável por receber as páginas ou outros objetos coletados e armazená-los em uma base de dados local e ainda por fazer a análise gramatical do texto, podendo tratar vários formatos como PDF, HTML, etc. O *servidor de nomes* atende a requisições DNS (Domain Name Service) dos coletores e mantém um *cache* de identificadores de nomes resolvidos. O *escalonador*, por sua vez, é responsável por decidir qual a próxima URL a ser coletada e coordenar as ações dos coletores.

O algoritmo básico executado por um robô é adquirir uma lista de URLs raízes como entrada e executar repetidamente os seguintes passos:

1. Remover a URL da lista de URLs;
2. Determinar o endereço IP dessa URL;
3. Carregar o documento correspondente;
4. Extrair os links contidos no documento;
5. Para cada link extraído, garantir que é uma URL absoluta e colocá-la na lista de URLs a serem carregadas, contanto que não já tenha sido carregada antes.

O Mercator [23] é um robô completamente escrito em linguagem de programação Java e tem como um de seus pontos fortes a extensibilidade. Por exemplo, novos módulos de protocolo podem ser providos para busca de documentos de acordo com os diferentes tipos de protocolos de rede, ou novos módulos de processamento podem ser providos para processar documentos carregados de diferentes formas personalizadas. Adicionalmente, o Mercator pode ser facilmente reconfigurado para utilizar diferentes versões da maioria dos seus componentes principais.

Os robôs atualmente utilizados, em geral, são capazes apenas de rastrear o conteúdo público da Web, ou seja, as páginas alcançáveis apenas através de links entre elas. No entanto, existe uma grande parte da Web que não pode ser alcançada desta forma, sendo necessário, por exemplo, requisições de formulários pelo usuário para que estas informações sejam extraídas de banco de dados. Raghavan e Garcia-Molina [24] propõem um modelo genérico para rastrear essa porção oculta da Web que, segundo estimativas recentes à publicação do artigo, representava cerca de 500 vezes o tamanho da Web pública. No trabalho, é abordada uma nova Técnica de Extração de Informação Baseada em Layout (do inglês, LITE - *Layout-based Information Extraction*), bem como a demonstração do seu uso para extração de informação semântica de formulários de busca e páginas de resposta.

2.2.2. Indexação

Os mecanismos de indexação nos sistemas de busca para Web utilizam técnicas similares às dos sistemas tradicionais de RI, em geral baseadas em arquivos invertidos. No Google [9][20], a função de indexação é executada por dois componentes de sua arquitetura, o indexador e o ordenador. Este indexador realiza diversas funções, como a leitura do repositório, a descompressão dos documentos e a realização do *parsing* destes documentos. Desta forma, cada documento é convertido em um conjunto de ocorrências chamadas de *hits*. Estes *hits* guardam informações acerca do posicionamento da palavra dentro do documento, além de algumas informações de apresentação como tamanho da fonte, utilização de negrito, etc. Além destas funções, o indexador ainda realiza o *parsing*, extraindo das páginas e armazenando em arquivos do sistema informações sobre âncoras e links. Estes arquivos resultantes contêm informações suficientes para determinar a estrutura de interligação das páginas, bem como o texto que as relaciona (texto das âncoras).

A Figura 2.6 mostra a estrutura de indexação utilizada pelo Google. Neste exemplo, o índice é formado segundo as WORDIDs, que são identificadores dos termos contidos nos documentos. Cada WORDID aponta para um conjunto de documentos que contêm o termo. Além do DOCID, armazena-se um conjunto de *hits*, que guardam informações acerca do posicionamento da palavra dentro do documento, além de algumas informações de apresentação.

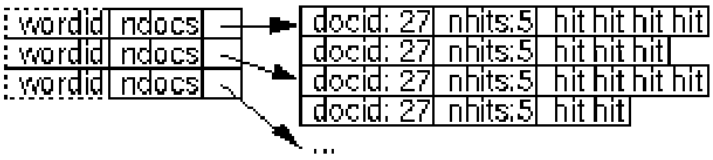


Figura 2.6 - Estrutura de indexação do Google [20].

2.2.3. Ranking de Relevância

O processo de elaboração do *ranking* de relevância em um sistema de busca para Web pode diferenciar-se do modelo adotado em sistemas tradicionais de RI, onde foca-se na

similaridade entre a consulta e os documentos. Como exemplo, tem-se o algoritmo de *ranking* utilizado pelo Google. Com o objetivo de mensurar a relevância das páginas da Web, Page et al [9] criaram o algoritmo PageRank. Segundo os autores, a Web pode ser representada como um grafo direcionado, onde os vértices são as páginas, as arestas os links entre elas e as direções o sentido de navegação determinado pelos links. A partir deste conceito, a grosso modo, uma página é dita mais importante quando possui uma quantidade maior de *backlinks* (*links* que apontam para esta página). Nota-se que isto é uma analogia a uma votação (cada *backlink* é um voto para a página). No entanto, existem diferenças nas importâncias dos votos, ou seja, um voto de uma página classificada como mais importante pode valer mais do que diversos votos de páginas pouco relevantes. Desta forma, temos que uma página pode ser mais relevante que outra mesmo tendo menos *backlinks*.

Além de algoritmos como o PageRank, outras características peculiares da Web podem ser utilizadas para construção do *ranking*, como por exemplo o uso de textos contidos nas âncoras, pois, frequentemente, estas fornecem uma descrição mais precisa das páginas Web para as quais apontam do que as próprias páginas. Outro exemplo pode ser a utilização dos textos contidos em *tags* especiais, como a *tag* “alt”. Ainda, a utilização de alguns detalhes visuais da apresentação, como o tamanho das fontes, pode ser útil. Palavras escritas em fontes maiores ou com negrito, por exemplo, podem ter um peso maior que outras palavras.

2.3. Recuperação de Informação Geográfica

A introdução da capacidade de processamento geográfico adiciona novas dificuldades e desafios aos processos de organização e recuperação de informação, exigindo a adaptação dos principais componentes de um sistema de RI. Nas seções a seguir são apresentadas as principais características de um sistema de GIR, bem como discutidas algumas soluções propostas para o comportamento dos principais processos relacionados.

2.3.1. Identificação de Características Geográficas

Uma tarefa importante em um sistema de GIR é de inferir de forma automatizada localizações geográficas associadas a documentos Web obtidos através de um robô. Alguns sistemas comerciais [3][4] oferecem busca através de localizações, porém baseando-se em serviços de diretórios. Apesar desta ser uma modalidade de GIR que possui dados bastante precisos, este tipo de sistema demanda bastante esforço para o cadastramento das páginas em suas respectivas classes, além de geralmente se limitarem a páginas de empresas. Alguns autores propõem a utilização de metadados específicos através de *tags* especiais para explicitar as informações espaciais dos documentos, como as descritas por [25][26] e pelo padrão ISO/TC 19115. No entanto, existem diversas dificuldades relacionadas à aplicação deste tipo de técnica, como por exemplo, o fato de os autores necessitarem que os sistemas de busca avaliem estes metadados como critérios de busca para que possam colocá-los em seus documentos, enquanto que os sistemas não consideram este tipo de informação porque ainda não há documentos contendo esses metadados. Outro problema é a existência de *webmasters* maliciosos que podem manipular estas informações com interesses próprios.

No entanto, ignorando a existência de metadados para este fim, existem ainda diversas maneiras de deduzir informações geográficas com base, por exemplo, no conteúdo das páginas e na estrutura de links da Web. Segundo Markowetz et al [28][29], este processo pode ser dividido em duas etapas: extração e mapeamento. Na primeira, são identificados os elementos que são utilizados para referenciar localidades geográficas, como por exemplo, nomes de lugares e códigos postais; na segunda, associa-se cada referência detectada a uma localidade geográfica válida. McCurley [27] discute sobre o *geocoding*, um processo de atribuição de localizações geográficas a páginas Web, com base em descrições textuais. McCurley [27] apresenta diversos termos importantes encontrados em páginas que podem ser utilizados para inferir localizações, como códigos postais, nomes de cidades e números de telefones; entretanto, não se discute maiores detalhes sobre técnicas de extração e de eliminação de ambiguidades.

Markowetz et al [28][29] utilizam o conceito de *geocoding* dividindo-o em três etapas: *geo extraction* (extração de termos candidatos do conteúdo e das URLs das páginas), *geo matching* (identificação dos termos extraídos e eliminação de ambiguidades) e *geo*

propagation (análise da estrutura topológica dos links para refinamento dos mapeamentos); um determinado documento pode ser associado a uma ou mais localizações, e este conjunto de localizações são chamados de *footprints* geográficos da página. No restante deste trabalho é utilizado, por vezes, apenas o termo *footprint* para referir-se ao *footprint* espacial de documentos e consultas.

Buyukkokten et al [30] e Ding et al [33] apresentam-se conceitos para escopo geográfico. No primeiro, o escopo geográfico é atribuído a uma página com base em dados coletados a partir do endereço IP dos servidores hospedeiros e representado por uma coordenada geográfica referente ao centróide da região resultante do processo de análise da página. No segundo, é apresentado um modelo de atribuição de escopos a partir dos conteúdos dos documentos e da estrutura de links da web, utilizando-se uma hierarquia cidade-estado-país das regiões administrativas dos Estados Unidos. Discute-se ainda sobre a implementação do modelo em um protótipo desenvolvido, chamado GeoSearch. Em Markowetz et al [32], relata-se que há bastante relevância nos dados da seção *admin-c* (seção que contém dados de contato do administrador do domínio) do *whois*; no entanto, segundo os autores, outras partes não possuem tanta importância, visto que estão relacionadas com a localização dos servidores e que muitas companhias pequenas ou mesmo indivíduos terceirizam o serviço de hospedagem, fazendo com que muitas vezes o local onde está armazenado o site não tenha nenhuma relação com o local de sua criação, ou mesmo com o seu conteúdo.

No protótipo apresentado por Gravano [2], utiliza-se *geocoding* automático com base nas idéias descritas por Ding et al [33]. Em Markowetz et al [28][29], observam-se diversas similaridades com o trabalho de Gravano [2] e Ding et al [33], diferenciando-se em alguns pontos como: (i) a granularidade das informações tratadas - os primeiros consideram simples páginas em suas análises, enquanto os demais fazem análises mais generalizadas, considerando apenas sites; (ii) O conjunto de sites investigados - os primeiros se restringem ao domínio “.de” e os demais ao domínio “.edu” e, em cada um dos casos, beneficiam-se de algumas peculiaridades destes domínios, como por exemplo, do fato do *whois* do domínio “.edu” prover, particularmente, boas estimativas sobre a localização do Web site e, ainda, da existência de uma lei alemã que exige que todo site do domínio “.de” possua alguma página, a

menos de dois cliques da página inicial, contendo os dados completos de contato do proprietário [28][29].

Várias técnicas para extração de características e classificação automática de documentos já foram propostas, entretanto, a obtenção de características geográficas de páginas Web envolve algumas dificuldades adicionais a estes tipos de processos. Dentre estas dificuldades, tem-se a ocorrência de ambiguidades (e.g. vários lugares com mesmo nome, vários nomes para o mesmo lugar, coisas com nomes de lugar, etc). Outra questão importante é que, por exemplo, pode-se ter uma determinada página que se refere a um determinado local que contém ou está contido no local requerido na consulta do usuário, fazendo com que seja necessário o conhecimento sobre a toponímia das diferentes regiões utilizadas pelo sistema. Para auxiliar o processo de extração de características geográficas dos documentos, faz-se necessário a consulta a mecanismos externos, onde um dos mais significativos são os *gazetteers* e *thesauri*. Pode-se citar como exemplo o TGN [34], um thesaurus utilizado em diversos projetos atualmente em desenvolvimento. O TGN mantém em sua base informações globais sobre nomes históricos de lugares (visto que o nome dos lugares pode ser alterados com o tempo), informações sobre hierarquias administrativas, coordenadas geográficas (em geral centróides), dentre outras informações úteis.

No processo de eliminação de ambiguidades proposto por Martins et al [35], todas as referências detectadas no texto são associadas a um peso. Ao fim do processo, seleciona-se uma única referência, a de maior peso associado, para representar o escopo geográfico do documento; ou nenhuma, se todos os pesos atribuídos forem inferiores a um limiar pré-estabelecido. Os valores atribuídos às entidades podem ser propagados a outras através de relacionamentos ontológicos entre estas, com a aplicação de métodos de inferência de modelos gráficos probabilísticos. Entretanto, o trabalho não apresenta maiores detalhes sobre os métodos aplicados.

Yi Li et al [36] descrevem um método elaborado para eliminação de ambiguidade, denominado de Resolução de Toponímia. Neste, atribui-se um valor de probabilidade a cada localidade possível para um dado nome ambíguo. Estas localidades são consultadas em um *gazetteer*. Nos experimentos descritos, foi utilizado o TGN. O valor inicial é dado segundo informações sobre a localidade recuperadas do TGN, como por exemplo, se esta é uma capital ou um local não habitado. Em seguida, outras heurísticas são utilizadas para incrementar os

valores iniciais: (i) ocorrência de localidades espacialmente relacionadas cujas referências estejam próximas no texto; (ii) estatísticas populacionais; (iii) termos geográficos (e.g. “country”) precedendo ou sucedendo a referência analisada. Nenhum dos locais é descartado completamente, podendo ser associado ao documento com um valor muito baixo de probabilidade. Não são revelados detalhes acerca dos métodos utilizados, como por exemplo, a que distância, no texto, as referências podem influenciar umas às outras; ou o peso de cada uma dessas variáveis no cálculo do valor final de probabilidade.

Volz et al [37] propõem um método para eliminação de ambiguidades baseado em uma ontologia alimentada a partir de fontes de dados externas. Inicialmente, as referências candidatas são associadas a um peso, calculado com base em propriedades da referência obtidas de um *gazetteer*, de forma análoga ao proposto para os valores iniciais de probabilidade em Yi Li et al [36]. Estes pesos podem assumir valores negativos. Em seguida, são analisados os termos textualmente vizinhos (com distância –5 à +5) aos termos que representam referências candidatas, em busca de palavras que possam definir a classe da localidade (e.g., cidade, país, montanha) e de referências correlacionadas geograficamente. As referências candidatas com peso final negativo são eliminadas e as demais são classificadas segundo um valor resultante da multiplicação do peso pelo número de ocorrências. Por fim, seleciona-se uma única referência como a mais provável para representar o escopo geográfico do documento.

Silva et al [38] discutem sobre técnicas de extração de características geográficas de grandes coleções de documentos Web, cujo método envolve o reconhecimento de referências geográficas no texto utilizadas para então atribuí-lo um escopo geográfico através de um algoritmo chamado de *GraphRank* [39], inspirado no *PageRank* [9]. Martins et al [39] mostram ainda três conjuntos de heurísticas utilizadas no processo de georreferenciamento de páginas Web, sobre a manipulação (i) de referências explícitas de conceitos geográficos em páginas Web, (ii) do ambiente Web e seus *hiperlinks*, (iii) das referências geográficas.

2.3.2. Modelagem do Escopo Geográfico

Uma vez identificadas as localidades referenciadas pelo documento Web, tem início o processo de modelagem do escopo geográfico do documento. Neste, são identificados os

lugares pelos quais o documento será indexado. O escopo geográfico pode ser simples ou múltiplo. No primeiro caso, este contém apenas uma única localidade associada, que muitas vezes é uma generalização das localidades identificadas no processo anterior; no segundo, um documento é associado a várias localidades, que podem ser as mesmas que foram identificadas ou outras relacionadas. As propostas encontradas na literatura variam ainda pela forma de representar o escopo geográfico (e.g., conjunto de centróides de cada localidade).

Silva et al [38] afirmam que o escopo de uma página da Web é associado apenas a uma localidade, podendo esta ser, por exemplo, uma generalização de outras localidades a ela relacionadas. Markowetz et al [28][29][40] sugerem que um documento possa ter escopo múltiplo, ou seja, que possa ser associado a várias localidades que, em muitos casos, possuem área não contínuas; nesta abordagem, o escopo é representado por um *footprint* espacial onde, para cada localização que contida neste *footprint*, associa-se um número inteiro representando o grau de certeza de que aquela página efetivamente contém informação relevante para tal localidade.

Amitay et al [41] propõem um modelo de atribuição de escopos geográficos múltiplos a documentos da Web, baseado em relacionamentos hierárquicos do tipo “parte de” (*part-of)*. Nesta abordagem, pode ser atribuída ao escopo geográfico uma localidade para a qual o documento não possui referência. Utiliza-se uma hierarquia de cidade-estado-país-continente extraído de um *gazetteer*. Para cada localidade pertencente ao escopo (referenciada ou não pelo documento), um algoritmo calcula um valor que representa a importância da localidade no documento. Esse valor é utilizado no processo de geração do *ranking* de relevância geográfico.

Zhisheng Li et al [42] apresentam um sistema de GIR que utiliza locais implícitos para prover ganho de desempenho. Locais implícitos são os ancestrais dos locais explícitos encontrados nos documentos (por exemplo, "América do Norte" é um local implícito para "Canadá" e para "Estados Unidos"). Propõem utilizar o algoritmo de Wang et al [43] para modelar o escopo geográfico das páginas Web e, em seguida, adicionar os ancestrais das localidades inicialmente detectadas como locais implícitos, no entanto com valores de relevância menores.

Yi Li et al [36] propõem um sistema de GIR dividido em 4 etapas: reconhecimento e classificação de entidades nomeadas (do inglês, NERC); resolução probabilística de toponímia

(do inglês, TR); indexação espacial; e recuperação. Nessa abordagem, atribui-se, durante o processo de eliminação de ambiguidades existentes em TR, um valor a cada localidade candidata. Este valor é calculado, dentre outros fatores, pela existência de relações hierárquicas entre as localidades encontradas no texto. Este modelo hierárquico, baseado no TGN, é utilizado também nas demais etapas descritas. O sistema proposto suporta a expansão do documento (ou indexação redundante), similar ao de Zhisheng Li et al [42], e a expansão de consultas. Entretanto, com a utilização de índices redundantes, relatam dificuldade para se atribuir diferentes pesos aos termos expandidos e para se realizar alguns tipos de consultas, como por exemplo, as baseadas em distância. Com isso, preferem resolver tais questões com a aplicação da expansão de consultas utilizando a estrutura hierárquica.

Overell et al [44] descrevem experimentos com um aplicativo de GIR, chamado Forostar [45]. Neste, as entidades extraídas do texto, classificadas como localidades, passam por um processo de eliminação de ambiguidades baseado em heurísticas, que as associa ao ID de uma localidade única do TGN. O sistema é implementado sobre o Apache Lucene (lucene.apache.org). Com isso, dois campos no índice do Lucene são utilizados para armazenar as informações geográficas sobre o documento: um contendo uma coordenada geográfica obtida do TGN e outro contendo uma string de representação do documento, no formato: id_ancestral_n\..\id_ancestral_1\id_localidade. Esta string é resultante de um processo de expansão de referências similar ao de Yi Li et al [36], uma vez que não atribui valores de relevância para localidades expandidas. O processo de busca é baseado na comparação de strings com mesma estrutura das utilizadas para indexação. Por exemplo, seja 123 o id para uma localidade *l* indexada por 789\456\123. Assim, a consulta “localidades em *l*” seria descrita por 789\456\123*, enquanto a expressão 789\456* representaria a consulta “localidades próximas à *l*”.

2.3.3. Indexação e Consultas Espaço-textuais

Após o processo de extração de características e adição de escopo geográfico às páginas, estas precisam ser indexadas de forma a promover a recuperação eficiente das informações armazenadas. A utilização de índices textuais tanto para dados textuais quanto

para dados espaciais significa empregar a mesma estrutura de arquivo invertido utilizado nos sistemas de busca tradicionais, onde é necessário que o argumento utilizado na consulta do usuário para referenciar o espaço geográfico coincida exatamente com o nome utilizado nos documentos armazenados. No entanto, sabe-se que, na prática, isto não acontece em todos os casos.

Uma alternativa é a expansão de consultas através de relacionamentos espaciais com a localidade especificada pelo usuário, por exemplo com regiões próximas. Supondo uma localidade A, próxima a B e C e uma consulta "carro AND A", a nova consulta gerada poderia ser "carro AND (A OR B OR C) ", no caso de uma heurística de expansão por proximidade. Outras heurísticas podem ser, por exemplo, a utilização de sinônimos para os nomes de lugares ou a adição de regiões contidas na região especificada. No entanto, este tipo de prática pode ocasionar o crescimento excessivo da consulta, por exemplo, quando a consulta original referencia uma região muito grande e tenta-se realizar a expansão incluindo todos os lugares que estão contidos dentro desta.

Propõe-se em [40][46][47][48] a utilização, além dos índices textuais, de índices espaciais organizados utilizando os *footprints* dos documentos armazenados. A utilização dos índices espaciais requer a geração do *footprint* da consulta que, quando utilizado para acessar o índice espacial, pode servir para o mesmo propósito que a expansão de consultas discuta anteriormente, podendo entretanto ser capaz de realizar consultas que poderiam ser impraticáveis devido ao problema da proliferação de termos durante a expansão.

Martins [48] apresenta um *survey* sobre estruturas de indexação conhecidas, propondo que estas sejam combinadas para a construção de um sistema de GIR eficiente. Diversos índices multidimensionais foram propostos para manipulação de dados espaciais, dentre eles *R-trees, grids, quad-trees, k-d-tree*, dentre outros [49]. Entretanto, o método de indexação espacial mais popular é o *R-tree*, uma árvore derivada de uma *B-tree*, dividindo o espaço em retângulos organizados hierarquicamente, podendo estar sobrepostos, conforme ilustrado na Figura 2.7. Com o objetivo de prover maior eficiência, o algoritmo de inserção da *R-tree* procura minimizar a quantidade de sobreposições e a área total dos nós usando várias heurísticas. Segundo Martins [48], um conjunto de heurísticas *R*-tree* [31], tem sido empiricamente validado como razoavelmente eficiente para coleções aleatórias de retângulos.

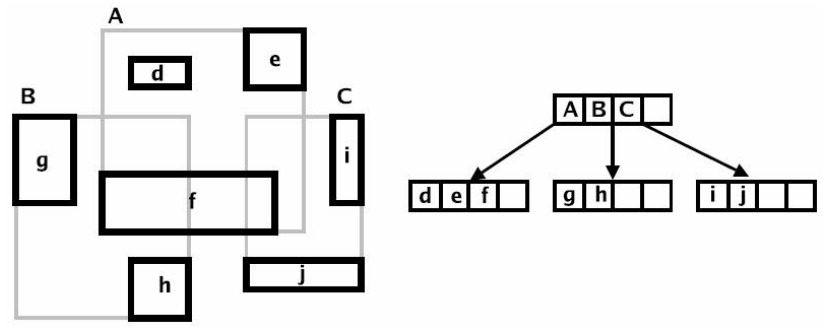


Figura 2.7 – Retângulos organizados hierarquicamente em uma R-tree [24].

Três abordagens para utilização de índices espaciais são mostradas por Vaid et al [47], comparado-as entre si e com índices puramente textuais (PT). A primeira, chamada de "Índice de Prioridade Espacial" (ST), prioriza a dimensão espacial no processo de indexação e acesso ao índice. A segunda, nomeada de "Índice Espaço-textual de Prioridade Textual" (TS), funciona de maneira inversa, priorizando a dimensão textual. Já na terceira, identificada como "Índice Textual com Processamento Espacial Posterior" (T), as dimensões textuais e espaciais são tratadas separadamente e, em seguida, é realizada a interseção dos resultados. Através de experimentos realizados, os autores concluem que o esquema T ocasiona menor custo de armazenamento e possui uma degradação muito pequena no tempo de consulta em relação ao esquema PT.

Uma discussão sobre o problema do processamento de consultas em sistemas de recuperação de informação geográfica é proposta por Chen et al [40]. No trabalho, são descritas algumas técnicas de indexação e alguns algoritmos para processamento de consultas. Dentre os algoritmos apresentados, parte deles é existente e outros são contribuições inovadoras, como os algoritmos *k-Sweep*, *Tile Index* e *Space-Filling Inverted Index*. Os algoritmos desenvolvidos foram integrados a um processador de consultas pré-existente de alta performance para sistemas de busca escaláveis. Todos os algoritmos apresentados foram comparados entre si. Os experimentos foram realizados com uma grande quantidade de documentos e consultas, ambos reais. Como resultado, observou-se que, em muitos casos, o processamento de consultas geográficas pode obter um desempenho com nível de eficiência muito parecido com o de consultas puramente textuais.

2.3.4. Ranking de Relevância

A elaboração do *ranking* de relevância é um dos principais processos em um sistema de busca, por vezes o responsável pelo seu sucesso, visto que está diretamente relacionado ao interesse do usuário. Em sistemas tradicionais, o *ranking* pode ser preparado utilizando-se diversas técnicas, a exemplo das medições de similaridade entre a consulta e os documentos utilizando o modelo espaço-vetor, discutido anteriormente neste trabalho. Um dos métodos mais populares atualmente é o PageRank [9], que utiliza a estrutura de links da Internet para elaboração do *ranking*.

Em motores de busca geográficos o problema torna-se um pouco mais complexo, visto que deve-se considerar a dimensão espacial para elaboração do *ranking*, além da utilização das técnicas tradicionais para a dimensão textual. No protótipo de Jones et al [46], emprega-se o algoritmo BM25 para a relevância textual e, para relevância espacial, utiliza-se (i) a distância entre o *footprint* da consulta e o *footprint* do documento e (ii) a diferença angular de direções cardinais no caso de consultas qualificadas direcionalmente. Neste, o componente responsável pela elaboração do *ranking* acessa uma ontologia geográfica para a obtenção do *footprint* dos lugares comparados com os *footprints* das consultas, bem como para aquisição de dados acerca dos relacionamentos espaciais do lugar, como exemplo, se este contém ou sobrepõe outros lugares. Jones et al [50] propõem a combinação da distância euclidiana entre centróides de lugares com distâncias hierárquicas, com o objetivo de criar uma distância espacial híbrida a ser utilizada para construção do *ranking* de relevância dos objetos recuperados.

Markowetz et al [28][29][40] adotam em sua implementação o modelo de dados raster, representando o *footprint* espacial dos documentos em uma estrutura de dados bitmap. Com isto, sugerem que a relevância geográfica seja dada pelo volume da interseção entre os *footprints* espaciais da consulta e do documento. Markowetz et al [51] afirmam que a ordem não depende apenas da relevância do assunto (expresso pelo por palavras-chave), mas também pela proximidade geográfica (expressa por nomes de lugares). Para os autores, dependendo da consulta a ser realizada, um critério pode ser mais importante que o outro. Com isto, propõe-se a utilização de um parâmetro de balanceamento que, dependendo do valor ajustado, ordenam-se os documentos recuperados priorizando um ou outro critério.

2.3.5. Interface com o Usuário

Em sistemas de busca tradicionais, a interface em geral é bastante simples, limitando-se a um espaço para informar as palavras-chave e ao resultado apresentado como uma lista de itens. Em sistemas de busca geográficos, é necessária a utilização de interfaces um pouco mais elaboradas, visto que o usuário necessita, além de informar as palavras-chave, escolher as regiões de interesse, bem como especificar possíveis relacionamentos espaciais.

Em [46], apresenta-se um protótipo que provê uma interface multi-modo, onde o usuário dispõe de um módulo para entrada de textos estruturados (contendo as palavras-chave, um nome de lugar e um relacionamento espacial para o lugar), um módulo para textos livres e um mapa para visualização das informações espaciais. Neste protótipo, o usuário pode escolher entre diversas alternativas oferecidas pelo sistema caso sua consulta inicial gere ambiguidades. O *footprint* gerado para a consulta pode ser visualizado no mapa antes desta ser definitivamente efetuada e o resultado da busca, por sua vez, pode ser visualizado tanto na lista textual quanto através de símbolos pintados no mapa.

2.3.6. Ontologias e a GIR

A necessidade de recuperação de informação eficiente apoiada por conhecimentos acerca de um domínio específico impulsionou o interesse pelo desenvolvimento de ontologias que modelem diversos conceitos associados. Nos sistemas de GIR, a utilização de ontologias pode ser de grande importância para representação de características geográficas dos documentos.

O protótipo apresentado por Jones et al [46][50] sugere a existência de um componente ontológico primário – uma ontologia de lugar – que provê a modelagem da terminologia e da estrutura do espaço geográfico. Esta ontologia tem papel fundamental, por exemplo, na interpretação das consultas dos usuários, na formulação de consultas do sistema, na geração dos índices espaciais, na elaboração do *ranking* de relevância e na extração de metadados. Nesta abordagem, além da ontologia espacial, sugere-se a manutenção de ontologias de domínios específicos, como por exemplo, ontologias turísticas.

Chaves et al [52] sugerem que os conteúdos dos documentos sejam convertidos em representações RDF e sejam armazenados em bases de dados XML. Recomendam-se ainda que o conhecimento geográfico seja atribuído aos documentos como recursos RDF adicionais. Neste trabalho, apresenta-se o GKB, um repositório baseado em um meta-modelo independente de domínio para integração de conhecimento geográfico coletado de diversas fontes.

2.4. Considerações Finais

Neste capítulo, foram apresentados os principais conceitos relacionados à recuperação de informação clássica e geográfica. Na seção 2.3, foram relatadas algumas das contribuições mais importantes disponíveis na literatura para cada uma das sub-áreas da GIR. Alguns destes trabalhos serão referenciados novamente na seção 4.8, onde é realizada uma análise comparativa destes com as propostas apresentadas no presente trabalho. No próximo capítulo, descrevem-se as pesquisas no campo da GIR que possuem maior quantidade de informações disponíveis na literatura.

Capítulo 3

Trabalhos Relacionados

As pesquisas no campo da GIR ainda são recentes, sendo observado o surgimento de alguns sistemas resultantes de iniciativas comerciais e acadêmicas. Grandes empresas como Google, Yahoo e Microsoft têm investido no desenvolvimento de aplicações que fazem uso do geoprocessamento, o que inclui a adaptação de seus motores de busca para o tratamento de informações geográficas, a exemplo do Google Local (local.google.com) e Yahoo Local (local.yahoo.com). Ainda no âmbito comercial, outros sistemas menos conhecidos têm surgido; porém, em geral, estão limitados a uma certa região geográfica. Como exemplo, citam-se o www.search.ch, cujo escopo está limitado à suíça; e o www.umkreisfinder.de, limitado às localidades alemãs.

O que se observa em comum entre as iniciativas comerciais é a ausência de documentação acerca de detalhes sobre seu funcionamento interno (e.g., algoritmos utilizados), sendo a documentação existente limitada a informações sobre a interação com o usuário, de maneira similar ao que acontece com os sistemas de busca tradicionais, devido à forte concorrência existente entre as empresas envolvidas. Existem alguns sistemas que se caracterizam como um sistema de GIR apenas por alguns aspectos bastante limitados, e não serão considerados neste trabalho. Possivelmente existe também uma grande quantidade de sistemas ainda desconhecidos dispersos pela Internet.

Aplicações como Google Local e Yahoo Local são baseadas em serviços de diretório, e em geral limitadas a sites de empresas. Estes possuem como fontes de informação, por exemplo, as páginas amarelas de listas telefônicas, de onde obtêm dados como endereço, telefones e URL de empresas. Assim, as páginas da web retornadas no resultado das buscas são aquelas contidas dentro do domínio de uma empresa cadastrada, ou então alguma outra

página da web contendo o endereço de uma destas. Porém, diferentemente das listas telefônicas, as empresas cadastradas não precisam pagar pelo serviço (apesar de também haver exposição de anúncios pagos) e, ainda, é possível solicitar a inclusão de uma empresa nova, caso esta ainda não esteja cadastrada. Em ambos os sistemas citados, o operador espacial utilizado no processo de busca é o "perto de", com base em uma única localidade especificada pelo usuário. Apesar desta ser uma importante proposta para busca geográfica, esta se diferencia consideravelmente do modelo de obtenção e busca de informações geográficas discutido neste trabalho. Por esta razão, e pela ausência de documentação relacionada, tais iniciativas não serão discutidas neste capítulo.

No meio acadêmico, alguns protótipos de sistemas de GIR têm sido apresentados, resultantes de esforços de diversos projetos de pesquisa. Algumas destas pesquisas se destacam por reunirem uma quantidade maior de documentação disponível na literatura, sendo possível conhecê-las em maiores detalhes. Portanto, neste capítulo, descrevem-se as principais características identificadas nestas pesquisas, comparando-as quanto às técnicas e modelos propostos.

3.1. GeoSearch

O GeoSearch - A Geographically-Aware Search Engine [30][33][53], é um dos mais antigos projetos ligados à GIR, e teve um importante papel para impulsionar as pesquisas na área. No entanto, o projeto encontra-se atualmente desativado, tendo seu último trabalho publicado no ano de 2003. Desenvolvido por um grupo de pesquisadores da universidade de Columbia e Stanford, tem como objetivo a criação de um protótipo de um sistema de GIR com suporte a localidades pertencentes às regiões administrativas dos Estados Unidos, contidos em uma hierarquia de cidade-estado-país.

O GeoSearch indexa 300 jornais americanos com publicação na Web. O sistema calcula o escopo geográfico de uma página com base em seu conteúdo, nos dados do endereçamento IP dos servidores hospedeiros e na distribuição dos links que apontam para a página. Os dados dos administradores dos domínios são coletados a partir de três bases de dados acessadas localmente, obtidas na Internet, uma vez que ferramentas como *whois*

possuem tempo de resposta insuficiente para prover a escalabilidade necessária em um processo de crawling na web. Cada página analisada possui um único escopo geográfico, que pode ser uma generalização de várias localidades de menor nível hierárquico. Por exemplo, em uma página contendo associação a várias cidades em um mesmo estado, o escopo geográfico pode ser generalizado como sendo o estado.

A interface Web do motor de busca, desenvolvida em linguagem de programação Java, contém um mapa dos Estados Unidos, onde podem ser visualizadas algumas respostas do sistema, bem como um conjunto de campos textuais utilizados para compor as requisições ao mesmo. Uma consulta é composta por uma lista de palavras-chave e por um código postal representando a localidade de interesse do usuário. Para uma dada consulta, os artigos jornalísticos correspondentes são recuperados utilizando-se um motor de busca textual chamado Swish (www.vic.com/swish/swish.html). Em seguida, são eliminados os documentos cujos escopos não estão relacionados ao código postal especificado. As páginas restantes são então reordenadas através de um novo valor de ordenação calculado, resultante da combinação do valor gerado pelo Swish e do valor relacionado ao escopo geográfico da página.

É possível ainda fornecer uma determinada URL ao sistema e visualizar graficamente as localidades de interesse pela página por esta endereçada. Um conjunto de sites do domínio ".edu" são analisados de modo a se inferir o contexto geográfico desses a partir dos dados do administrador destes domínios. Com isto, uma vez informada uma URL ao sistema, este consulta um serviço disponibilizado pelo Google para saber quais URLs (dentre o conjunto de URLs analisado) possuem links para a página especificada. Em seguida, as localidades relacionadas às URLs recuperadas são exibidas no mapa através de marcas circulares, onde o raio do círculo varia de acordo com a quantidade de links que apontam para a página.

Observa-se que a utilização de informações do administrador do domínio para inferir o escopo geográfico dos documentos não funciona bem para todos os domínios, uma vez que grande parte das páginas está hospedada em servidores localizados em regiões distintas do produtor da informação e, algumas vezes, esta informação também não está associada com o local onde foi produzida. Deste modo, o uso de tais informações é bastante limitado, fazendo com que os autores se restringissem a utilização de domínios ".edu", onde há maior relação entre tais localidades.

Conforme mencionado, utilizam-se técnicas de generalização para simplificar o escopo atribuído às páginas. Porém, note que esta prática pode fazer com que informações sejam perdidas, podendo ocasionar redução da qualidade do resultado da busca, de acordo com a pesquisa realizada pelo usuário. No que se refere à interface com o usuário, verifica-se ainda que a possibilidade de se especificar a dimensão geográfica da consulta de forma visual, através de seleções no mapa, proporcionaria ao usuário uma melhor interação com o sistema.

3.2. SPIRIT

O projeto SPIRIT (*Spatially-Aware Information Retrieval on the Internet*) [1] é dos principais projetos da atualidade com foco em recuperação de informação. O projeto foi financiado pelo *EC Fifth Framework Programme* e tem recebido colaboração de seis parceiros europeus. O projeto relaciona diversas técnicas ligadas à recuperação de informação geográfica: interface multi-modo que provê tanto a entrada de dados textuais como a interação com um mapa que relaciona o contexto geográfico dos documentos recuperados; ontologias que modelam o espaço geográfico; expansão de consultas e *ranking* de relevância baseados em ontologias geográficas; índices espaciais para as coleções de documentos e um mecanismo de aprendizado para extração de contexto geográfico a partir de documentos para geração de metadados espaciais. Uma visão geral do projeto é apresentada por Jones et al [46], abordando aspectos sobre sua arquitetura e sobre o uso de ontologias e índices.

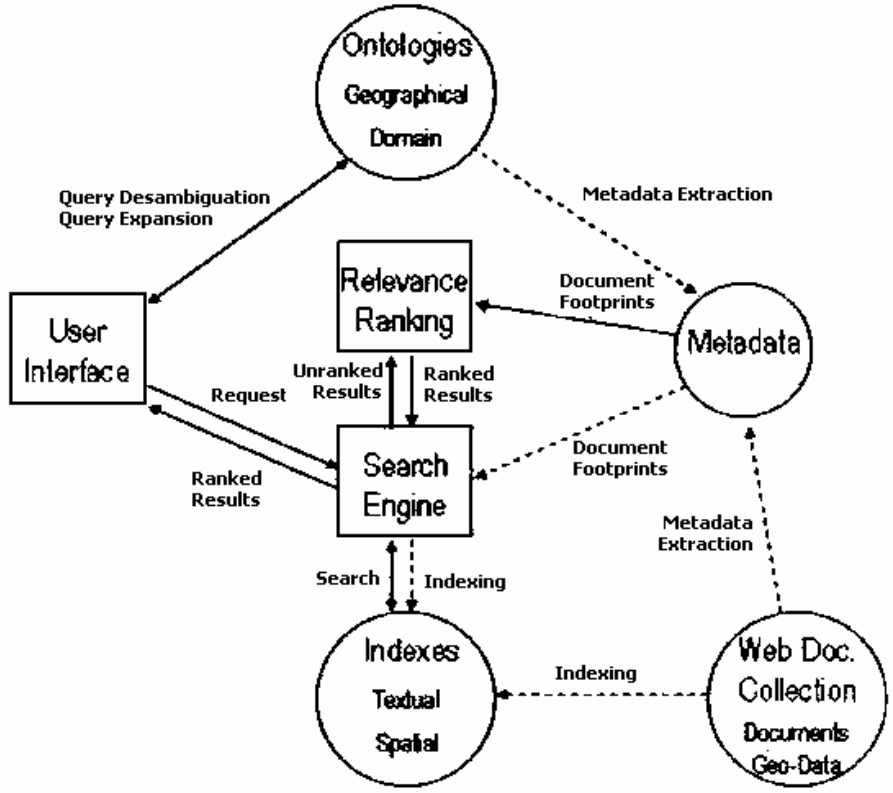


Figura 3.1 – Arquitetura do SPIRIT Search Engine [46].

Na Figura 3.1, observa-se a arquitetura do SPIRIT: uma coleção de documentos Web (*Web Documents Collection*) contendo dados espaciais e não-espaciais é armazenada. Para deduzir as características espaciais desses documentos, esses passam por um processo de extração de metadados, gerando os seus *footprints*. Os documentos são indexados (*Indexes*) por suas dimensões textuais e espaciais. Para realização do processo de extração, ontologias (*Ontologies*) podem ser acessadas. Estas ontologias, por sua vez, são importantes para a eliminação de ambiguidades e para a expansão das consultas oriundas da interface com o usuário (*User Interface*). O motor de busca (*Search Engine*) recebe a consulta da interface e, através dos índices, recupera os documentos que a satisfaz, ordena-os segundo sua relevância utilizando o componente de ordenação (*Relevance Ranking*) e retorna-os para a interface.

Jones et al [50][54] apresentam uma ontologia de lugar, que associa dados limitados de coordenadas com relacionamentos espaciais entre lugares. Expõe-se ainda o conceito de medida híbrida de distância espacial, que combina distâncias euclidianas entre centróides de lugares com medidas de distâncias hierárquicas, com possibilidade de agregação com

distâncias temáticas baseadas em classificações semânticas, com o objetivo de criar uma medida semântica integrada, utilizada para ordenar os objetos recuperados de acordo com seu grau de relevância.

Neste sentido, o objetivo é associar um nome de lugar específico a nomes de lugares que sejam similares ou equivalentes geograficamente. Para isto, assume-se que um lugar pode ser um fenômeno geográfico (como um rio, uma montanha), um conceito cultural, ou mesmo administrativo, como é o caso de cidades e estados. Estes fenômenos e conceitos determinam os tipos de lugares. A ontologia de lugar foi utilizada em um sistema chamado OASIS (Ontologically-Augmented Spatial Information System), que tem sido utilizado para manter informações culturais sobre arqueologia. Na ontologia proposta (vide Figura 3.2), observa-se, dentre outros elementos, um elemento *Lugar (Place)* que possui auto-relacionamentos *encontra (meet)*, *sobrepõe (overlap)* e *parte de* (*partOf)*. Um elemento *Artefato (Artifact)* se relaciona com lugar através de *encontrado_em (found_at)* e *feito_em (made_at)*. O elemento *Lugar* possui ainda um relacionamento com uma *Localização (location*; simplificada a um centróide), um *Tipo Atual (current place type)* e um *Histórico de Tipos (historical place type)*, um *Nome Preferencial (Preferred Term)* e um conjunto de *Nomes Alternativos (Non Preferred Term)*.

No que diz respeito às medidas de similaridade, não é de interesse comparar lugares por formas e estruturas e sim, por um conjunto de relacionamentos espaciais como contém, está contido, toca, dentre outros. Neste projeto, utilizou-se uma combinação entre a distância euclidiana e distância hierárquica (uma medida que leva em consideração relacionamentos genéricos do tipo *partOf*, que podem ser interpretados espacialmente como sendo *inside* ou *overlap*).

O modelo ontológico de Jones et al [50][54] é mostrado em maiores detalhes em Smart et al [55] e Jones et al [56]. Apresentam-se definições de lugar (i.e. lugar absoluto, que contém informações sobre seu nome, tipo e localização; e lugar relativo, que descreve o lugar baseado em relacionamentos com outros lugares). Apresentam-se ainda definições para *footprints* de consultas e de documentos. Discute-se sobre a aplicação da ontologia proposta na arquitetura de um sistema de recuperação de informação geográfica. Apresentam-se também modelos e exemplos de possíveis consultas ao sistema, bem como considerações de projeto relacionadas a diversos elementos, como nomes e tipos de lugares, endereços, localizações, relações

espaciais, sistema de coordenadas, tempo, linguagem, generalizações, dentre outros. Discutem-se maiores detalhes sobre a ontologia de lugar, como os requisitos de linguagem para sua descrição, exemplos de codificação em XML, questões de manutenção e operadores de acesso.

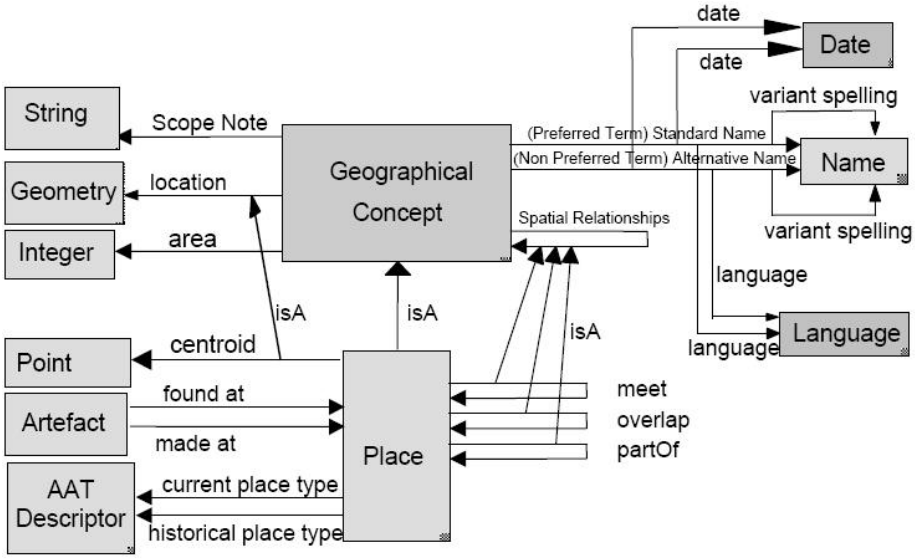


Figura 3.2 – Modelo de lugar no sistema OASIS [10].

Vaid et al [47] realizam uma análise sobre indexação espaço-textual para busca geográfica na Web. São exibidas três abordagens para construção de índices, comparado-as entre si e com o índice puramente textual (PT), baseado no modelo tradicional de arquivo invertido.

- Índice de Prioridade Espacial (ST): Prioriza a dimensão espacial no processo de indexação e acesso ao índice. Neste modelo, o espaço geográfico correspondente à cobertura imposta pelos nomes de lugares encontrados nos documentos é dividido em células regulares, formando um *grid*. Para cada uma destas células, é construído um arquivo invertido que funciona da mesma maneira que o índice PT, contendo os documentos cujos *footprints* intersectam a célula espacial.
- Índice Espaço-textual de Prioridade Textual (TS): Funciona de maneira inversa ao anterior, priorizando a dimensão textual. Neste, o modelo é PT é modificado

de forma que, para cada lista de documentos associada a um determinado termo do índice, esta seja separada em grupos organizados espacialmente. Ou seja, para cada termo do índice, os documentos as referências aos documentos serão organizadas da forma [$Célula_1$(Lista de $Documentos_1$); $Célula_2$(Lista de $Documentos_2$); $Célula_n$ (Lista de $Documentos_n$)], onde cada uma destas listas de documentos contém os documentos cujos *footprints* intersectam a célula associada.

- Índice Textual com Processamento Espacial Posterior (T): Nesta abordagem, utiliza-se um esquema PT para selecionar um conjunto A de documentos contendo os termos não geográficos e, separadamente, um índice espacial para recuperar um conjunto B de documentos cujos *footprints* intersectam o *footprint* da consulta. O resultado final é então obtido através da interseção de A e B.

Por meio dos experimentos realizados, os autores concluem que o esquema T ocasiona menor custo de armazenamento e possui uma degradação muito pequena no tempo de consulta em relação ao esquema PT.

Atualmente o projeto SPIRIT encontra-se em atividade, com trabalhos recentemente publicados. O grupo possui papel significativo no progresso das pesquisas em GIR, não apenas com os resultados efetivos de suas pesquisas, mas também através da promoção de importantes eventos na área, com o objetivo de promover a difusão do conhecimento e o intercâmbio entre pesquisadores. Os trabalhos publicados abrangem diversas sub-áreas da GIR, no entanto, observa-se a ausência de detalhes sobre o processo de extração de características geográficas, no que diz respeito às heurísticas utilizadas e aos algoritmos aplicados. As atividades planejadas concentram-se no aprimoramento da geo-ontologia desenvolvida e no aperfeiçoamento das técnicas de indexação espaço-textual. Tais atividades incluem a integração de múltiplas fontes de dados para construção de uma geo-ontologia multinacional, a representação dos *footprints* geográficos em várias escalas, a representação de nomes de lugares imprecisos e a realização de experimentos com maiores coleções de documentos georreferenciados.

3.3. Geographic Search Engine for Germany

O projeto descrito por Markowetz et al [28][29] tem como objetivo desenvolver um protótipo com as idéias contidas em [27][32][33], adicionando algumas contribuições. O protótipo proposto é baseado em uma coleção de 30 milhões de páginas do domínio ".de" capturadas utilizando um robô pré-existente. Para simplificação dos experimentos, os dados utilizados e as análises realizadas estão relacionados apenas a Alemanha. Para tais experimentos, conta-se com bancos de dados contendo informações, por exemplo, sobre mapeamentos entre cidades e códigos de área utilizados pela telefonia ou mesmo entre cidades e códigos postais. Utiliza-se o conceito de *geocoding* dividindo-o em três etapas: *geo extraction*, *geo matching* e *geo propagation*. Para cada localização contida no *footprint* atribuído à página, associa-se um número inteiro representando o grau de certeza de que aquela página efetivamente contém informação relevante para a localidade especificada. A seguir, relacionam-se mais alguns detalhes acerca das três etapas mencionadas:

Geo Extraction: Esta etapa consiste em identificar, em um determinado documento, os termos candidatos a serem mapeados a uma localização geográfica, como nomes de cidades, números de telefone e códigos postais. Além do conteúdo da página, analisa-se a URL. Para os nomes de cidades, um cadastro manual foi feito com o objetivo de separar *termos fortes* (palavras que são utilizadas exclusivamente para referenciar cidades) de *termos fracos* (palavras que podem referenciar outras coisas além de cidades). A idéia é detectar primeiramente os termos fortes para então detectar os fracos segundo algumas restrições, como, por exemplo, a distância que um determinado termo fraco está de um termo forte. Outros tipos de termos são tratados de forma especial no processo de extração, como os "termos assassinos" (termos que fazem com que um termo forte seja ignorado caso esteja a uma certa distância dele), os "termos validadores" (termos que fazem com que um termo forte só não seja ignorado se estiver a uma certa distância dele), os "termos assassinos gerais" (para os quais todos os termos fortes a uma certa distância deles, são ignorados), dentre outros.

Geo Matching: Após a etapa anterior, os documentos são reduzidos a conjuntos de termos candidatos. Esta etapa possui como entrada estes conjuntos, onde cada termo é mapeado a uma cidade, e então a uma localização geográfica. Para este processo, algumas hipóteses são tratadas como verdadeiras, como por exemplo, a de que o autor do documento

refere-se a apenas um lugar, quando vários lugares possuem o mesmo nome, ou mesmo a de que o autor referencia a cidade com maior área, dentre as várias com mesmo nome. O grau de certeza do mapeamento entre um nome de cidade e um conjunto de termos pode ser medido de várias formas. No projeto, utiliza-se um valor para representar este grau de certeza. Para esta etapa, os autores propõem um algoritmo bem específico para o contexto alemão, chamado BBFirst, chamado desta forma por extrair primeiramente a melhor das maiores cidades.

Geo Propagation: Diversas páginas que fazem parte de um contexto geográfico podem não conter referências explícitas a lugares em seu conteúdo. Desta forma, esta terceira etapa tem como objetivo propagar *footprints* entre páginas, de acordo com a estrutura topológica dos links entre elas. Segundo os autores, páginas com links ou co-links entre si podem herdar propriedades geográficas. Por exemplo, em um site de uma empresa, pode ser que as informações geográficas explícitas estejam restritas à página de contatos da empresa, no entanto, é possível que todas as demais páginas do site façam parte do mesmo contexto geográfico. No modelo proposto, o grau de relacionamento entre as páginas é mensurado segundo um fator α, onde $0 < \alpha < 1$. Por exemplo, para páginas contidas no mesmo diretório, o valor de α é maior que para páginas apenas no mesmo site.

Em seus trabalhos, Markowetz discute ainda sobre técnicas de indexação e otimização de consultas. Chen et al [40] apresentam alguns algoritmos de indexação simples, já conhecidos na literatura:

- *Text-First Baseline:* Primeiramente, buscam-se os resultados através dos termos indexados textualmente em um índice em arquivo invertido. De posse dos documentos recuperados, recuperam-se todos os *footprints* associados a cada documento. Os *footprints* são indexadas pelos IDs dos documentos.
- *Geo-First Base Line:* Nesta abordagem, recuperam-se inicialmente os documentos cujos *footprints* possuem interseção não vazia com o *footprint* da consulta. Os footprints são aproximados utilizando MBR (Minimum Bounding Rectangle) e indexados pelos IDs dos documentos. Em seguida, os IDs recuperados são ordenados e filtrados utilizando o índice em arquivo invertido.

Argumenta-se que o primeiro possui algumas vantagens em relação ao segundo, entretanto que existem casos em que o segundo pode ter um desempenho melhor. Chen et al [40] discutem sobre alternativas para a disposição dos dados em disco de forma a diminuir o

número de acessos e, dessa forma, otimizar o processamento de consultas. Em seguida, propõem três novos algoritmos, para os quais recomendam a utilização de *toeprints* ao invés de *footprints* e a utilização de *curvas space-filling* [57][58] para *layout* dos dados em disco. *Toeprints* nada mais são do que subconjuntos disjuntos de *footprints*, resultantes da divisão de um *footprint* MBR (Figura 3.3a) em porções menores (Figura 3.3b) de forma a diminuir a quantidade de espaços vazios dentro de um MBR. Para representações espaciais, foram utilizados *grids* regulares de 1024x1024 células, onde cada célula representa uma área de 700x1000 metros. Os novos algoritmos propostos são:

K-Sweep: este algoritmo tem como objetivo recuperar todos os dados de *toeprints* através de um número fixo k de varreduras no disco. Para cada célula no *grid*, armazena-se uma lista de *m* intervalos de IDs de *toeprints*, indicando que todo *toeprint* que intersecta esta célula possui o ID dentro dos intervalos especificados. Inicialmente, o sistema busca os intervalos para todas as células que fazem interseção com o *footprint* de uma certa consulta requisitada. Em seguida, computa $k \geq m$ intervalos maiores (*sweeps*), que representa a união dos intervalos destas células. Uma limitação apontada para este algoritmo é que ele recupera estes dados completos dos *toeprints* sem antes filtrar pelos termos da consulta (consulta textual).

Tile Index: este algoritmo visa eliminar as limitações do anterior. Nele, armazena-se para cada célula do *grid* uma lista completa de todos os IDs de *toeprints* que intersectam a célula. Além disso, existe uma tabela que mapeia IDs de *toeprints* em suas localizações no disco e seus IDs de documentos. Desta forma, é possível que os IDs de documentos sejam filtrados utilizando o arquivo invertido antes dos dados dos *toeprints* serem recuperados.

Space-Filling Inverted Index: Esta abordagem visa alterar o arquivo invertido propriamente dito, segundo critérios espaciais. A proposta é utilizar IDs de *toeprints* como IDs de documentos. Caso um documento possua mais de um *toeprint*, criam-se várias entradas para este documento através de IDs diferentes durante a criação do arquivo invertido. Esta idéia pode ser combinada com qualquer um dos dois algoritmos descritos anteriormente.

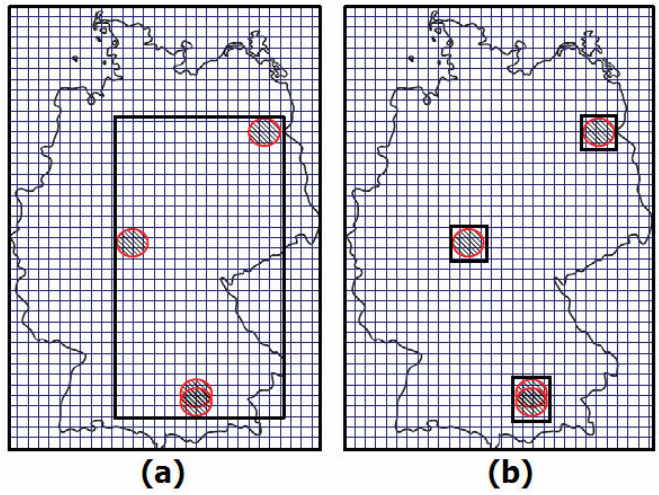


Figura 3.3 – Em (a), footprint MBR; em (b), toeprints [58].

Através de experimentos utilizando diversas configurações para os algoritmos, vários tamanhos de *footprints* de consultas e *grids* de diferentes resoluções, observou-se que, em relação aos algoritmos ingênuos, o tempo de consulta foi diminuído bastante em todos os algoritmos propostos e, dentre estes, notou-se uma melhora gradativa na seguinte seq uência: k-Sweep, Tile Index e Space-Filling Inverted Index. Com este último foi possível observar, em alguns casos, tempos de resposta em níveis equivalentes aos de consultas puramente textuais.

Outras contribuições importantes foram dadas por este grupo. Em Markowetz et al [32][51] foi introduzido o conceito de *locality*, que mede o quanto uma página está relacionada com sua vizinhança geográfica, sendo útil, por exemplo, para saber se uma página é de interesse global ou de apenas uma região restrita. Propõe-se um método de *ranking* que integre *page rank* com proximidade espacial, enfatizando-se que somente com a possibilidade de balanceamento dinâmico entre esses dois fatores é possível uma navegação flexível. Ainda argumenta-se em [32][51] sobre a construção de um robô espacialmente dependente e integrado a um *data warehouse*. A proposta seria o uso de robôs locais, restritos a sites de uma determinada região. O resultado da coleta destes robôs locais são armazenados separadamente em *data marts*.

O grupo apresenta contribuições importantes nas diversas sub-áreas da GIR, fornecendo em seus trabalhos detalhes sobre os algoritmos e técnicas utilizadas. Alguns procedimentos utilizados ainda possuem possibilidades de melhoria, como por exemplo, o

mecanismo de eliminação de ambiguidades que, dado duas localidades de mesmo nome, restringe-se a selecionar a de maior contingente populacional. Outros métodos desenvolvidos, como o de cálculo de confiança no processo de *Geo Extraction*, poderiam ser adaptados de forma a se viabilizar suas utilizações; por exemplo, com a aplicação de um procedimento automatizado para classificação dos nomes de lugar (e.g., fortes e fracos), uma vez que o processo atual depende de esforço manual para serem categorizados. Infelizmente, atualmente o projeto encontra-se descontinuado e, nos trabalhos publicados, relatam-se apenas o desenvolvimento do protótipo descrito, sem demais informações que evidenciem a conclusão de uma versão funcional.

3.4. GeoTumba!

O Tumba! (iniciais de Temos Um Motor de Busca Alternativo!), um projeto Desenvolvido pelo Grupo XLDB, do Laboratório de Sistemas Informáticos de Grande Escala (LASIGE) da Faculdade de Ciências da Universidade de Lisboa (FCUL), tem como um de seus objetivos o desenvolvimento de uma ferramenta de busca para a Web portuguesa empregando diversas técnicas desenvolvidas através de pesquisas realizados no projeto. O GeoTumba! é uma extensão do Tumba!, que visa adicionar ao sistema de busca uma contextualização geográfica, sendo resultante da integração de resultados com um outro projeto de pesquisa, o GREASE (Geographic Reasoning for Search Engines).

Em experimentos realizados, o grupo constatou uma ocorrência média de 2.2 referências por documento a algum dos 308 municípios portugueses em um total de 3.775.611 páginas analisadas. Além disso, cerca de 4% de um conjunto de consultas requisitadas ao *tumba!* possuíam referência a alguns desses municípios. Se a análise levasse em consideração outras diversas possibilidades de nomes de lugares além desses nomes de municípios, esta estatística aumentaria consideravelmente. Silva et al [38] apresentam um conjunto de heurísticas para detecção de referências geográficas em páginas da Web, dividido em três grupos. Abaixo, estão algumas das principais heurísticas apontadas em cada um dos grupos:

- Informações textuais em Páginas Web

- Se uma referência geográfica está presente em uma página, o escopo da página está relacionado àquela região;
- A importância de um termo para um documento pode variar de acordo com o lugar onde o termo aparece no documento e aumenta com o número de ocorrências do termo no documento;

- Hiperlinks e o ambiente Web
  - O escopo geográfico de um documento possui relação com a localização do servidor onde está hospedado;
  - Referências geográficas estendem sua influência além da página onde estas ocorrem, até mesmo para páginas de outros sites;
  - Páginas interligadas através de hiperlinks podem possuir escopos geográficos similares e os textos das âncoras podem conter informações importantes sobre o escopo da página de destino;
- Uso de referências geográficas:
  - Cada página contém apenas um escopo geográfico (podendo ser, por exemplo, uma generalização de escopos menores);
  - Se um documento referencia uma região, este pode estar relacionado a outras regiões espacialmente ligadas a esta, como por exemplo, regiões adjacentes, sub-regiões, etc;
  - Lugares maiores e/ou mais populosos devem ser escolhidos no processo de eliminação de ambiguidades;

O framework para mineração de dados da Web do tumba! é apresentado na Figura 3.4. Neste modelo, os dados são coletados e armazenados em uma Base XML (*XML Base*), com seus conteúdos convertidos em uma representação RDF. Estes documentos são chamados de Documentos Web Purificados ou PWD (*Purified Web Documents*). Em seguida, uma série de transformações incorporam nos PWDs conhecimentos adicionais sobre os documentos, tranformando-os em Documentos Web com Escopos Ampliados ou SAWD (*Scope Augmented Web Documents*). Os componentes responsáveis por tais transformações são chamados de amplificadores (*augmenters*), utilizados para um domínio específico. Para a atribuição de escopo geográfico aos documentos, utilizam-se dois amplificadores, um para reconhecimento de referências geográficas no texto e eliminação de ambiguidades e o outro para atribuição de

escopo utilizando as referências identificadas e um algoritmo chamado de *GraphRank* [39], similar ao *PageRank* [9].

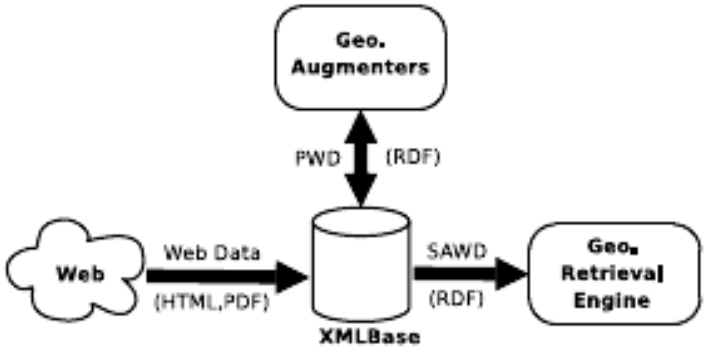


Figura 3.4 - Visão Geral do Projeto [27].

Um componente fundamental no processo descrito acima é o GKB (*Geographic Knowledge Base*) [52][51], um repositório baseado em um meta-modelo independente de domínio que contém o conhecimento geográfico utilizado pelo framework, através da integração de múltiplas fontes, provendo armazenamento sob um esquema único e ainda mecanismos para manter e exportar os conhecimentos adquiridos, dentre eles o conhecimento geográfico. A exportação se dá através da geração de ontologias OWL que, por sua vez, podem ser utilizadas por diversas aplicações da Web Semântica. O GKB inclui informações encontradas em *gazeteers*, relacionamentos entre *features* geográficas, dados demográficos, códigos postais, coordenadas geográficas, dentre outras. Possui algumas similaridades com o TGN [34], no entanto, possui ainda informações sobre o domínio da Internet e regras para manipulação de entidades geográficas e seus relacionamentos. Com estas regras é possível, por exemplo, descrever o conhecimento de que vários municípios portugueses possuem Web sites hospedados em domínios cujo nome contém os prefixos “cm-” ou “mun-” [38].

Atualmente, o projeto encontra-se em plena atividade, integrando pesquisadores de diversos níveis, desde alunos de graduação até pós-doutores. Como aspectos positivos desta pesquisa, citam-se a disponibilização de uma versão funcional do motor de busca para acesso público, e a iniciativa de prover uma versão dirigida a dispositivos móveis, o que é de grande utilidade na área, visto o relacionamento direto entre a mobilidade e o posicionamento

geográfico. Dentre os aperfeiçoamentos planejados, cita-se a realização de buscas geográficas com combinação de locais e com uso de relações espaciais mais complexas (e.g., “entre Porto e Espinho”, “fora de Lisboa”, “adjacente a Braga”, “ao norte de Évora”). Pretende-se ainda efetuar a desambiguação da pesquisa utilizando o histórico do utilizador, permitir o agrupamento dos resultados das pesquisas em clusters de âmbitos geográficos e classificar de forma mais exata os âmbitos geográficos das páginas Web.

3.5. Conclusão

Neste capítulo, foi possível conhecer alguns dos mais importantes projetos de pesquisa em GIR, selecionados com base na quantidade de documentação disponível na literatura. Para facilitar a comparação entre os projetos apresentados, foram selecionadas 27 características importantes de um sistema de GIR. Em seguida, relacionam-se estas características aos principais trabalhos descritos. Estas informações estão sumarizadas na Tabela 3.1, com os relacionamentos representados por uma marcação ●. A ausência desta marca significa que a característica não está presente ou que esta informação não foi encontrada na literatura.

Descrição	**Geo Search**	**SPIRIT**	**GSE for Germany**	**Geo Tumba**
Distingue entre escopo geográfico de uma página Web e locais de interesse por esta página.			●	
Considera fenômenos geográficos como lugares (e.g., um rio, uma montanha).		●		
Calcula valores de confiança para as referências geográficas detectadas no texto.			●	●
Considera alguns termos como especiais no processo de detecção de referências geográficas.	●	●	●	●
Analisa a formatação dos elementos em documentos HTML (e.g., uso de negrito, itálico, coloração e outras *tags* de destaque).				●

Descrição	Geo Search	SPIRIT	GSE for Germany	Geo Tumba
Analisa a grafia das referências no processo de detecção (e.g., uso de maiúsculas e abreviações nos nomes de lugares, forma como um cep está descrito, etc).				
Possui uma base de dados com valores pré-calculados de probabilidade de um determinado nome referir-se a algum lugar por este identificado.				
Para uma referência detectada, altera seu valor de confiança com base em relacionamentos espaciais com outras localidades diferentes também detectadas no texto.				
Utiliza informações contidas nas URLs.	●		●	●
Utiliza informações do Whois.	●		●	●
Utiliza estatística entre os termos e os textos (e.g. número de ocorrências, posicionamento no texto).	●			●
Permite associar escopos geográficos múltiplos aos documentos.		●		
Utiliza estatísticas sociais (e.g., contingente populacional) nos processos relacionados à atribuição de escopo geográfico e elaboração do ranking de relevância.		●	●	●
Utiliza ontologias para apoiar o processo de modelagem e manipulação de informações geográficas semânticas.		●		●
Analisa a estrutura de ligação entre as páginas da web (*links*).	●	●	●	●
Atribui escopo geográfico e elaboração do ranking de relevância com base no padrão de distribuição dos locais referenciados.				
Possui interface multi-modo.	●	●		●
Permite ao usuário balancear, em tempo de execução, entre a relevância geográfica e textual para elaboração do *ranking*.			●	
Utiliza técnicas de expansão de índice.				

Descrição	Geo Search	SPIRIT	GSE for Germany	Geo Tumba
Calcula valores de relevância para as localidades expandidas do índice.				
Disponibiliza três ou mais operadores espaciais para execução da busca.		●		●
Permite selecionar visualmente, como argumento espacial de busca, localidades pré-definidas e/ou especificar regiões por seleções livres.		●		●
Considera zonas temáticas como localidades válidas no processo de busca (e.g., região com maior índice de violência, região com maior renda per capita).				
Suporta recuperação de recursos multimídia (e.g., imagens, vídeos) com base em seus escopos geográficos.				
Possui versão para aplicativos móveis.				●

Tabela 3.1 – Características de um sistema de GIR verificadas nos trabalhos relacionados.

Após conhecer os trabalhos relacionados e suas características mais relevantes, apresentam-se no próximo capítulo as propostas do presente trabalho, bem como os detalhes sobre o funcionamento do protótipo desenvolvido para validação destas, em seus aspectos arquiteturais e comportamentais.

Capítulo 4

GeoSEn: um Motor de Busca com Enfoque Geográfico

Neste capítulo, apresenta-se o GeoSEn, um protótipo de um motor de busca para Web, com enfoque geográfico, que implementa os modelos e técnicas elaborados para construção de um sistema de GIR. Primeiramente, serão exibidos os principais componentes da arquitetura do sistema; em seguida, descrevem-se os métodos e técnicas elaborados para a construção dos mecanismos de detecção de referências geográficas, de indexação espaço-textual, de modelagem do escopo geográfico e de execução de buscas considerando as dimensões textual e espacial. Apresenta-se ainda uma interface multi-modo, contendo um mapa interativo para auxiliar a entrada de informações espaciais durante a interação com o usuário.

4.1. Arquitetura do Sistema

O GeoSEn tem como princípio a utilização de softwares livres. Estes são adotados, por exemplo, na plataforma operacional, nas ferramentas de programação e de gerência de projeto e nos servidores de aplicação e de gerenciamento de bancos de dados. Seguindo esse princípio e visando explorar ao máximo o reuso de software, foi realizado um trabalho investigativo com o objetivo de escolher um motor de busca e um robô de códigos fontes abertos a serem estendidos com as funcionalidades previstas para o GeoSEn. Dentre os nove robôs e onze motores de busca avaliados, o Nutch (lucene.apache.org/nutch/) foi selecionado em ambas as categorias, dentre outros motivos:

- por ser um projeto em contínuo desenvolvimento, assegurando a utilização de um *software* com maiores possibilidades de evolução;
- por possuir uma quantidade razoável de documentação, quando comparado aos demais, atenuando as dificuldades de desenvolvimento;
- por integrar em um único sistema tanto um robô quanto um motor de busca, minimizando os custos de integração desses módulos;
- por ser escrito sobre o Lucene que, por sua vez, vem sendo mencionado na literatura como um *framework* robusto de manipulação textual. Além disso, uma vez que se utiliza o índice textual no formato do Lucene, torna-se possível o uso ferramentas gratuitas já desenvolvidas para manipulação destes índices;
- pela extensibilidade do sistema, proporcionada por sua arquitetura orientada a *plugins*, facilitando a integração com os recursos do GeoSEn.

A escolha do SGBD PostgreSQL (www.postgresql.org), por sua vez, deu-se em virtude do elevado grau de maturidade do projeto e por possuir uma extensão espacial bastante robusta, o Postgis (www.postgis.org), que contém os recursos necessários para implementação das funções espaciais planejadas.

A Figura 4.1 exibe a arquitetura do GeoSEn. O sistema é desenvolvido sobre o Nutch, um motor de busca *Open Source* da Apache Software Fundation (www.apache.org). O Nutch, por sua vez, é desenvolvido sobre o Lucene, um *framework* também da Apache que fornece uma API para indexação textual e o núcleo (*core*) de um sistema de busca. Para o Lucene, não importa a origem dos dados, seu formato ou linguagem, ficando a cargo dos seus usuários a implementação dos módulos de recuperação dos documentos das fontes desejadas e de análise de tais documentos de forma a convertê-los para o formato de texto simples reconhecido pelo Lucene.

Desta forma, o Nutch adiciona ao Lucene, dentre outras funcionalidades, a capacidade de recuperar as páginas da Web (*crawling*) e de analisar os documentos (*parsing*) em diversos formatos, como por exemplo, HTML e PDF. Todos os referidos softwares são escritos em linguagem de programação Java.

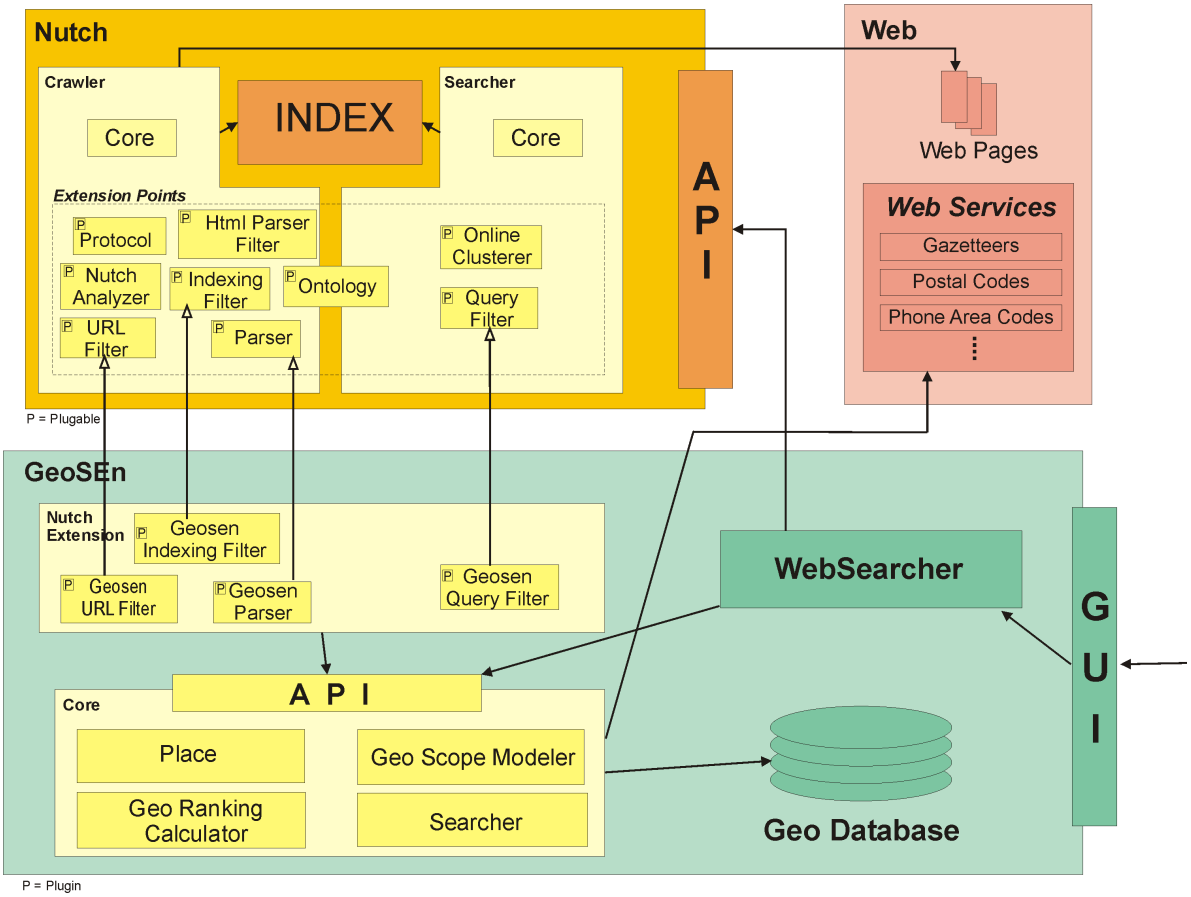


Figura 4.1 - Arquitetura do GeoSEn

A arquitetura do Nutch divide-se basicamente em duas partes: o *Crawler* e o *Searcher*. O *Crawler* é o robô responsável por coletar as páginas da Web, analisá-las e então indexá-las para futura recuperação. O índice utilizado é no mesmo formato de índice do Lucene, tornando-o compatível com as ferramentas e APIs disponíveis para análise e manipulação deste tipo de índice. Este robô mantém ainda uma estrutura de dados chamada WebDB para representar a estrutura de grafo da Web (onde as páginas são os vértices e os *links* as arestas) utilizada no processo de seleção das páginas a serem capturadas. O processo de *crawling* é dividido em dez etapas, que podem ser executadas automaticamente, através de um único comando, ou executadas individualmente, possibilitando maior controle pelo usuário administrador do sistema. O modo manual é mais utilizado nos processos de atualização e manutenção do índice.

O *Searcher* recebe e interpreta as consultas (*queries*) fornecidas pelos usuários, consulta o índice e retorna os resultados que satisfazem a consulta. O *Searcher* pode ser acessado de várias formas, como por exemplo, utilizando-se a Nutch API, que provê acesso diretamente ao Nutch através da linguagem de programação Java. Outra alternativa é utilizar a OpenSearch API (www.opensearch.org), uma extensão do RSS 2.0 para publicar resultados de motores de busca – indicado para clientes escritos em outras linguagens de programação.

O Nutch possui uma arquitetura orientada a *plugins* e oferece diversos pontos de extensão. Na Figura 4.1, uma área identificada por *Extension Points* destaca tais pontos que, por sua vez, são identificados na arquitetura como pontos plugáveis (*Plugable*). Dentre esses pontos, foram escolhidos os que seriam de interesse para o funcionamento do GeoSEn. Ilustra-se, na área correspondente ao GeoSEn, os plugins que foram implementados. São eles:

- Parser – Estes *plugins* realizam a leitura dos documentos capturados e extraem os dados a serem indexados. Os *plugins* dessa categoria são desenvolvidos se houver necessidade de extrair um novo tipo de conteúdo ou se for necessário coletar outros dados dos tipos de conteúdo já passíveis de extração. Para esta funcionalidade, o GeoSEn conta com o *plugin* Geosen Parser, responsável por detectar termos geográficos no processo de *parsing*, utilizados em seguida para análise do escopo geográfico dos documentos;
- Indexing Filter – Tais *plugins* permitem acrescentar novos campos ao índice do Nutch, possivelmente contendo novas características extraídas dos documentos provenientes do processo de *parsing*, as quais se deseje considerar durante a recuperação destes documentos através de consultas ao índice. Desta forma, o *plugin* Geosen Indexing Filter adiciona ao índice as informações sobre o escopo geográfico dos documentos, resultantes do processo de *parsing*;
- Query Filter - Permitem adicionar metadados às *queries*. O *plugin* Geosen Query Filter adiciona ao sistema a capacidade de interpretar as especificações de lugares nas consultas dos usuários, bem como de converter tais consultas para o formato utilizado para acesso ao índice. Assim, o modo como devem ser acessados os valores dos novos campos criados em um *plugin* do tipo Indexing Filter deve ser especificado em outro do tipo Query Filter. Portanto, o Geosen Query Filter tem

como função prover a capacidade de se consultar o índice considerando as informações sobre o escopo geográfico adicionadas pelo Geosen Indexing Filter.

- URL Filter - As implementações desse tipo de *plugin* possibilitam analisar as URLs acessadas pelo *web crawler*, sendo possível, inclusive, criar formas de especificar as que devem ser acessadas ou ignoradas. O *plugin* Geosen URL Filter oferece mecanismos para detectar referências geográficas nas URLs acessadas no processo de *crawling*.

Esses *plugins* comunicam-se com o núcleo do GeoSEn, que executa as regras de negócio relacionadas às funcionalidades dos *plugins*, bem como as demais funcionalidades do GeoSEn. Ou seja, os *plugins* funcionam apenas como uma ponte entre o Nutch e o GeoSEn, pois, de fato, as regras de negócio relacionadas às funcionalidades dos *plugins* estão implementadas no núcleo do GeoSEn. A comunicação entre o núcleo e os *plugins* se dá através do acesso a uma API disponibilizada pelo núcleo. Desta forma, retiram-se dos plugins as funcionalidades do GeoSEn, fazendo com que este esteja menos acoplado à arquitetura do Nutch. Isso torna possível a utilização das funcionalidades do GeoSEn por outros sistemas de busca com um menor esforço de adaptação, utilizando o acesso direto à API ou, como no caso do Nutch, com a criação de entidades intermediárias que se adequem à arquitetura do sistema e que acessem a API do GeoSEn.

Fazem parte do núcleo do GeoSEn os módulos de detecção de lugares (*Place Detector*), de modelagem de escopo geográfico (*Geo Scope Modeler*), de elaboração do *ranking* de relevância geográfica (*Geo Ranking Calculator*) e de execução de buscas considerando a perspectiva espacial (*Searcher*). Estes módulos utilizam uma base de dados contendo dados geográficos diversos (*Geo DataBase*), bem como serviços disponibilizados na Web (W*eb Services*), que provêem dados geográficos adicionais de interesse do sistema.

O acesso aos serviços externos para coleta dos dados necessários aos processos citados é realizado previamente à execução destes, em processos complementares executados periodicamente, sendo os dados recuperados replicados na base local. Essa replicação faz-se necessária por questões de desempenho, visto o número elevado de consultas que são necessárias para obtenção destes dados. Para se ter uma idéia, algumas destas consultas são realizadas para cada termo de cada documento analisado durante o *parsing*. Desta forma,

evitam-se perdas, por exemplo, com tempo de conexão e transmissão, com indisponibilidade do serviço, dentre outras.

Os dados contidos no *Geo DataBase* são coletados de diversas fontes. Os dados básicos e espaciais acerca das localidades reconhecidas pelo sistema foram obtidos do IBGE (Instituto Brasileiro de Geográfica e Estatística), bem como certos dados sociais e econômicos. Alguns dados complementares foram obtidos na Wikipedia (pt.wikipedia.org), como por exemplo, os gentílicos. Os códigos postais utilizados são os estabelecidos pelos Correios e os códigos de área telefônicos foram obtidos da Anatel.

O *Geo DataBase* é implementado utilizando o SGBD PostgreSQL, com sua extensão espacial Postgis e a biblioteca de indexação textual Tsearch2, utilizada principalmente para indexar os nomes de lugares. Por fim, há no GeoSEn um módulo chamado *Web Searcher*, que faz a ligação do sistema com o meio externo. É responsável por receber as consultas dos usuários através da interface gráfica Web (WGUI – *Web Graphic User Interface*), consultar os índices textuais e espaciais e então retornar os documentos que satisfazem à consulta em ambos os aspectos. Este módulo comunica-se com o núcleo do GeoSEn e com o Nutch através de suas respectivas APIs. As próximas seções deste capítulo explicam com maiores detalhes o funcionamento dos principais módulos do protótipo desenvolvido, a saber descreveremos os módulos de detecção de referências geográficas e eliminação de ambiguidades, de modelagem do escopo geográfico, de indexação e de busca.

4.2. Detecção de Referências Geográficas

Integrante do processo de *crawling*, o mecanismo de detecção de referências geográficas tem como objetivo identificar e extrair, nos documentos capturados pelo robô, o máximo de dados que possuam potencial de serem convertidos em informações geográficas, como por exemplo, nomes de lugares, códigos postais, códigos de área telefônicos, dentre outros. Identificados estes dados, tem início o subprocesso de conversão destes em localidades geográficas reconhecidas pelo sistema. Por exemplo, um código de área telefônico pode ser convertido em uma referência geográfica de um estado da federação. As referências geográficas podem ser detectadas no corpo do texto, no título da página (em casos de

documentos HTML) e na URL associada ao documento. As referências detectadas são filtradas por um processo de eliminação de ambiguidades e, em seguida, utilizadas na modelagem do escopo geográfico do documento, conforme está descrito nas seções subsequentes.

4.2.1. Confiança dos Termos Geográficos

Os termos geográficos detectados pelo *parser* são associados a um grau de confiança (CR – *Confidence Rate*), que é definido como:

> ***Confidence Rate (CR):*** *medida associada a uma referência geográfica detectada pelo parser, que representa a probabilidade desta ser ou não uma referência válida à localidade especificada.*

Definição 4.1 – Confidence Rate (CR)

O grau de confiança é o principal fator utilizado na eliminação de ambiguidades, assumindo valores no intervalo de 0 a 1. Estabeleceu-se um limiar, com valor 0,5, onde os termos com valor de CR inferior a este limite são ignorados. Assim, as referências detectadas que possuem baixa probabilidade de serem uma referência geográfica são ignoradas. Isto acontece, por exemplo, por existir em lugares com nome de coisas ou de pessoas, como em *Arame*, *Telha* e *Cláudia*, todos nomes de municípios brasileiros. Quando uma referência do tipo *nome de lugar* é detectada, e está relacionada ao problema de ambiguidade onde mais de um lugar é identificado pelo mesmo nome, como por exemplo, *London* no *Canadá* e *London* na *Inglaterra*, calcula-se um valor de CR para cada localidade possível, e então se seleciona aquela com maior valor.

O processo de detecção de referências geográficas analisa diversas características relacionadas às referências candidatas, com o objetivo de constituir o valor de CR com base no resultado destas análises. Tais características serão apresentadas mais adiante neste capítulo. Contudo, tem-se que o valor de CR é proveniente dos valores de entidades denominadas como

fatores de confiança (CF – *Confidence Factor*), que estão associadas à referência geográfica analisada. Este fator é definido como:

> ***Confidence Factor (CF):*** *medida associada a cada característica analisada durante o processo de detecção de uma referência geográfica. Cada CF possui um determinado peso no cálculo de CR. Portanto, o valor de CF, ponderado ao seu respectivo peso, representa o quão influente é essa característica na composição do valor de CR.*

Definição 4.2 – Confidence Factor (CF)

Os CFs utilizados e seus respectivos pesos podem variar de acordo com o local de detecção da referência (e.g., título da página, URL, corpo do texto), com o tipo de referência (e.g., nomes de lugar, CEP, telefone) e com o tipo de lugar (e.g., município, estado). Os CFs utilizados pelo GeoSEn estão relacionados abaixo e serão explicados em maiores detalhes adiante, bem como o cálculo do valor de CR.

- CF_{ST} – Analisa a ocorrência de termos especiais associados às referências geográficas;
- CF_{TS} – Considera probabilidades resultantes de consultas textuais;
- CF_{CROSS} – Analisa a ocorrência de referências cruzadas;
- CF_{FMT} – Avalia a sintaxe utilizada para descrever as referências geográficas;

O valor de cada CF, por sua vez, é obtido a partir dos valores de seu(s) modificador(es) de confiança (CM – *Confidence Modifier*). Define-se este modificador por:

> ***Confidence Modifier (CM):*** *medida mais detalhada utilizada na composição do valor de um CF. Em geral é utilizada quando uma determinada referência se relaciona repetidas vezes com uma característica do mesmo tipo, dentre as características avaliadas para formação do CF. Cada CF está associado a um ou mais CMs. Portanto, para um fator relacionado à característica X e identificado por* CF_X*, seus n modificadores são identificados por* CM_{X1}*,* CM_{X2}*, ... ,* CM_{Xn}*.*

Definição 4.3 – Confidence Modifier (CM)

Portanto, tem-se que uma referência pode estar associada a um ou mais fatores de confiança, enquanto cada fator pode estar associado a um ou mais modificadores. Os conceitos introduzidos nesta seção ficarão mais claros após conhecer as características avaliadas pelo sistema no processo de detecção de termos geográficos e como estas estão relacionadas aos seus modificadores e fatores. Tais elementos serão apresentados a seguir.

4.2.2. Reconhecimento de Termos Especiais

Uma das características analisadas durante a avaliação das referências geográficas candidatas é a ocorrência de termos especiais (ST - *Special Terms*), cujos são definidos por:

> ***Special Term (ST):*** *termo cuja ocorrência em um texto pode ser decorrente da presença de referências geográficas no mesmo texto*.

Definição 4.4 – Special Term (ST)

Exemplos de STs são: "em" (e.g. "em João Pessoa); "cidade" (e.g. "cidade de São Paulo"); “CEP” (e.g. “CEP: 58109-000”). Assim, o fator de confiança CF_{ST} de uma referência geográfica é modificado de acordo com a quantidade e o tipo dos STs a ela relacionados. Ou seja, cada termo especial relacionado a uma mesma referência representa um modificador do tipo CM_{ST}, cujos valores serão combinados para compor o valor de CF_{ST}. A Tabela 4.1 exibe os principais atributos de um ST:

Campo	Descrição
Termo	O termo especial
Tipo de referência geográfica	Quais tipos de referências geográficas que este pode influenciar (e.g., nomes de lugar, cep, telefone).
Tipo de lugar	Quais tipos de lugares este pode influenciar (e.g., município, estado).

Campo	Descrição
Distância Mínima (D_{MIN})	Distância mínima à referência associada para que tenha efeito sobre ela. Pode ser positiva ou negativa. Por exemplo, na expressão "em São Paulo" o termo "em" está a uma distância -1 da referência "São Paulo". Já na expressão "na cidade de Campina Grande" os termos "na" e "cidade" estão, respectivamente, a uma distância -3 e -2 da referência "Campina Grande".
Distância Máxima (D_{MAX})	Distância máxima à referência associada para que tenha efeito sobre ela. Pode ser positiva ou negativa. Deve ter o mesmo sinal da distância inicial.
Grau de Confiança Máximo (C_{MAX})	Grau máximo de confiança adicionado à referência pelo termo especial.

Tabela 4.1 - Atributos de um termo especial

O valor do modificador de confiança CM_{ST} é calculado de acordo com a Equação 4.1:

$$CM_{ST} = \frac{D_{MAX} - D + 1}{D_{MAX} - D_{MIN} + 1} \bullet C_{MAX}$$ Equação 4.1

Onde:

- *D é a distância calculada entre o termo e a referência;*
- $0 \leq CM_{ST} \leq 1$.

Assim, quanto mais próximo o termo especial está de sua referência associada, maior a sua influência no valor do modificador. Por exemplo, para um dado ST, se a distância mínima é igual a 2 e a distância máxima é igual a 4, há 3 possibilidades (quantidade de níveis) para que o ST possa adicionar confiança à referência, ou seja, se estiver a uma distância 2, 3 ou 4 da referência. Assim, se estiver a uma distância 2, adiciona confiança máxima à referência.

Estando a uma distância 3, adiciona 2/3 da confiança máxima. Por fim, estando a uma distância 4, adiciona 1/3 da confiança máxima.

Cada referência geográfica pode ser associada a um ou mais termos especiais. O valor final do fator de confiança CF_{ST} de uma referência geográfica é obtido conforme a Equação 4.2:

$$CF_{ST} = 0{,}5 + \sum_{i=1}^{n} \left(CM_{STi} \bullet 0{,}15\right), \text{ caso } n > 0; \qquad \text{Equação 4.2}$$

$$CF_{ST} = 0, \text{ caso contrário.}$$

Onde:

- *n é a quantidade de CM_{ST} associados à referência;*
- *$0 \leq CF_{ST} \leq 1$.*

Como valor máximo para CF_{ST} é 1, é alterado para este valor qualquer resultado maior. As constantes da Equação 4.2 foram obtidas empiricamente após avaliações experimentais com documentos contendo diversas situações de ocorrência de termos especiais. A primeira constante, igual a 0,5, faz com que a ocorrência de apenas um termo especial seja suficiente para que o valor de CF_{ST} seja maior que a metade de seu valor máximo. Já a segunda constante, igual a 0,15, faz com que a ocorrência de mais de 3 termos especiais com valor máximo de CM_{ST} ocasione a atribuição ao fator de seu valor máximo.

Por exemplo, suponha uma referência, a qual foram associadas três STs, com valores de CM_{ST} iguais a 0,80, 1,00 e 0,50, respectivamente. Desta forma, o valor calculado para o CF_{ST} desta referência seria 0,5 + (0,15 x 0,80) + (0,15 x 1,00) + (0,15 x 0,50), totalizando 0,845. Caso nenhum termo especial seja associado à referência, $CF_{ST} = 0$.

O detector de termos especiais do GeoSEn conta com 34 termos cadastrados, cujos atributos foram escolhidos de forma empírica com base em avaliações experimentais com dezenas de documentos contendo situações diversas de ocorrência destes termos. O ANEXO I lista os termos escolhidos com seus respectivos atributos.

4.2.3. Atribuição de Confiança a partir de Buscas Textuais

Visando resolver o problema da ambiguidade para os nomes de lugares que também são nomes de coisas, ou mesmo de pessoas, desenvolveu-se um mecanismo capaz de extrair características e associar graus de probabilidade com base no resultado da busca textual (TS – *Textual Search*) para o nome de um lugar. Durante a execução deste processo, cada nome de lugar cadastrado no sistema é analisado e o valor resultante é armazenado na base de dados. Deste modo, durante o processo de *parsing*, os valores previamente calculados são recuperados e representam o CM_{TS} da referência analisada. Este processo é definido como:

> ***Extração de Confiança de Buscas Textuais (TS):*** *procedimento que visa associar a um nome de lugar a probabilidade deste, isoladamente em um texto, se referir ou não à localidade por este identificada, baseando-se na análise do resultado fornecido por um motor de busca textual para o respectivo nome.*

Definição 4.5 – Extração de Confiança de Buscas Textuais (TS)

Ao se consultar um sistema de busca textual por um determinado nome, cada item do resultado contém um pequeno fragmento do conteúdo do documento onde há a ocorrência deste nome. No procedimento descrito, estes fragmentos são analisados em busca de certas palavras-chave que podem indicar se estes se referem ou não a uma localidade geográfica. Por exemplo, o fragmento "...Campina Grande é a segunda cidade mais populosa do estado da Paraíba...", extraído de um dos resultados para a busca textual por "campina grande", contém algumas palavras-chave (e.g., "cidade", "populosa") que podem indicar que este fragmento está descrevendo algo acerca de um município. Assim, os valores de CM_{TS} são obtidos com base na quantidade de palavras-chave verificadas no resultado da busca. Em seguida, estes valores são normalizados, de forma a se estabelecerem em uma faixa entre 0 e 1.

Para ilustrar a eficácia deste mecanismo, observe os seguintes nomes ambíguos de municípios brasileiros: *Ângulo*, *Arame*, *Bugre*, *Caracol*, *Coluna*, *Passagem*, *Saúde*, e *Sério*. Para estes (e outros), nenhum dos itens do resultado do motor de busca textual se refere aos municípios relacionados. Assim, assume-se que, na ocorrência desses nomes em documentos Web, a probabilidade (isoladamente) destes se referirem aos municípios relacionados é mínima. Estes são exemplos de nomes com valor zero para CM_{TS}. Por outro lado, alguns exemplos de municípios com valores altos de CM_{TS} são: *Monte Alegre de Sergipe*, *Nova Mamoré*, *Ouro Verde de Goiás*, *Potirendaba* e *São José dos Quatro Marcos*. Assim, é possível perceber que os nomes com valores mínimos são bastante comuns, sendo muito difícil afirmar, isoladamente, se possuem no texto a função de identificar um lugar. Por outro lado, têm-se valores altos para nomes bastante peculiares de municípios, onde a probabilidade é bastante alta de estes nomes se referirem, no texto, aos municípios por eles identificados.

Uma vez que os sistemas de busca exibem seus resultados segundo critérios de relevância dos documentos, é possível realizar a análise utilizando como amostra apenas os primeiros resultados retornados. O GeoSEn utiliza os 10 primeiros resultados. Observa-se que este é um processo de retroalimentação, onde se utilizam os resultados obtidos pelo processo de busca textual para retroalimentar o sistema, enriquecendo o mecanismo de busca espacial. Este é um exemplo de característica verificada pelo *parser* onde o fator CF_{TS} está sempre associado a um único modificador. Portanto, nestes casos, CF_{TS} e CM_{TS} possuem sempre o mesmo valor.

4.2.4. Referências Cruzadas

Outra característica avaliada pelo sistema é a existência de referências cruzadas, definida como:

***Referência Cruzada**: referência geográfica encontrada no texto, a qual possui determinados relacionamentos espaciais em relação a uma outra referência analisada. Seja $L(R_X)$ o lugar referenciado pela referência geográfica R_X. Assim, para uma referência analisada R_A, uma referência R_B é considerada referência cruzada de R_A se algum dos itens a seguir for verdadeiro:*

- *$L(R_B)$ contém $L(R_A)$;*
- *$L(R_B)$ está contido em $L(R_A)$;*
- *$L(R_A)$ e $L(R_B)$ estão em um mesmo nível hierárquico e existe um lugar L_X, tal que L_X seja do nível imediatamente superior a $L(R_A)$ e $L(R_B)$ e L_X contenha $L(R_A)$ e $L(R_B)$.*

Definição 4.6 – Referência Cruzada

Para ilustrar, considere o estado da *Paraíba*. Para uma referência geográfica R_P a este estado, são consideradas referências cruzadas a R_P as referências R_1, R_2, R_3,..., R_N; onde $L(R_1)$, $L(R_2)$, $L(R_3)$, ..., $L(R_N)$ são:

- a região que contém espacialmente $L(R_P)$, ou seja, a região Nordeste;
- todas as mesorregiões, microrregiões e municípios contidos espacialmente em $L(R_P)$;
- demais estados da *região nordeste*, onde todos estão no mesmo nível hierárquico de $L(R_P)$ e igualmente contidos na *região nordeste*;

Desse modo, cada referência cruzada verificada para uma determinada referência geográfica é representada por um modificador do tipo CM_{CROSS}. O valor deste modificador varia de acordo com a distância, no texto, entre as referências analisadas e a distância hierárquica entre os lugares por elas referenciados. Este valor é calculado utilizando a Equação 4.3:

$$CM_{CROSS} = \frac{0,5}{N \bullet D}$$ Equação 4.3

Onde:

- $0 \leq CF_{CROSS} \leq 1$;
- *N é o nível hierárquico;*
- *D é o fator de distância textual:*
  - *D = 1, se 1 ≤ distância textual ≤ 10*
  - *D = 2, se 10 < distância textual ≤ 20*
  - *D = 3, se 20 < distância textual*

O valor final do fator de confiança CF_{CROSS}, por sua vez, é obtido de acordo com a Equação 4.4:

$$CF_{CROSS} = 0{,}5 + \sum_{i=1}^{n} CM_{CROSSi} \text{ , caso } n > 0; \qquad \text{Equação 4.4}$$

$$CF_{CROSS} = 0, \text{ caso contrário.}$$

Onde:

- *n é a quantidade de* CM_{CROSS} *associados à referência;*
- $0 \leq CF_{CROSS} \leq 1$.

Como o valor máximo para CF_{CROSS} é 1, é alterado para este valor qualquer resultado maior. A constante da Equação 4.4 foi obtida de forma análoga à da Equação 4.3.

Observou-se que a identificação de referências cruzadas desempenha um papel importante na resolução de ambiguidades, principalmente no caso onde mais de um lugar possui o mesmo nome. Por exemplo, considere um texto onde foi detectada uma referência ao município de *Atalaia*. No entanto, no *Brasil*, existem dois municípios com este nome, localizados, respectivamente, nos estados de *Alagoas* e do *Paraná*. Diante disso, o sistema criará internamente duas referências, onde apenas uma será selecionada, após serem comparados os valores de confiança calculados para ambas. Porém, considere ainda que o texto mencionado referencia também o estado de *Alagoas*. Assim, o valor de CM_{CROSS} calculado para a referência *Atalaia / Alagoas* será maior que o calculado para a *Atalaia /*

Paraná. Esta diferença é considerada na eliminação de ambiguidade, o que pode influenciar na seleção da referência *Atalaia / Alagoas* em detrimento à outra referência analisada.

4.2.5. Formato das Referências

Alguns tipos de referências geográficas podem ser identificados no texto em diversos padrões de escrita. Para estes, são associados diferentes graus de confiança para cada um dos padrões possíveis. Por exemplo, uma referência do tipo CEP pode ser representada pelas sequências alfanuméricas “58.109-000”, “58109-000”, dentre outras. Esse fator de confiança é identificado por CF_{FMT} e seu(s) modificador(es) por CM_{FMT}.

No reconhecimento de referências do tipo *nome de lugar*, utilizam-se dois modificadores específicos. O primeiro tem a função de avaliar se um nome de lugar detectado está escrito com as letras iniciais em maiúsculas. Esta é uma técnica simples, porém, de grande utilidade na resolução de problemas de ambiguidade, uma vez que nomes de lugares, segundo a norma da língua portuguesa, devem conter as letras iniciais em maiúscula (exceto as palavras consideradas *Place Stop Words*, *que* são ignoradas pelo *parser* na busca por nome de lugares), mesmo quando não estão em início de sentença. Os valores possíveis para o modificador são 0 ou 1. Se algum dos termos que compõem o nome não tiver letra inicial maiúscula, o valor do modificador é zero. Caso todos os termos atendam a tal requisito, atribui-se o valor 1 ao modificador. Já o segundo CM_{FMT} para nomes de lugares é utilizado para mensurar o quão abreviado está o nome detectado. Quanto menos abreviado, mais confiável deve ser a referência. O cálculo desta taxa se dá pelo quociente entre o número de termos abreviados e o número total de termos que formam o nome do lugar (incluindo as *Place Stop Words*), assumindo valores entre 0 e 1.

Para ilustrar a importância da detecção de nomes abreviados, observe os resultados do Google para algumas abreviações do município *Cabo de Santo Agostinho*: para a expressão "cabo de s agostinho" foram encontradas por volta de 749 ocorrências e, para "cabo de santo a", aproximadamente 453 ocorrências. Para exemplificar utilizando cidades maiores, observa-se cerca de 38.100 ocorrências para a pesquisa por "R. de Janeiro", representando possivelmente documentos que referenciam a cidade e/ou o estado do *Rio de Janeiro*. São reconhecidos os nomes de lugares que contenham ao menos um termo não abreviado e onde

os termos abreviados sejam formados por apenas uma letra seguida ou não do caractere ponto. Durante a execução do *parser*, os termos do texto passíveis de compor um nome de lugar (contendo apenas caracteres alfabéticos) são subetidos para consulta à base de dados, que contém os nomes de lugares indexados textualmente. Deste modo, são retornados os nomes de lugares contendo esse termo em qualquer posição. Em seguida, de acordo com a quantidade de termos do nome candidato e com o posicionamento do termo analisado, compara-se cada possibilidade de abreviação gerada pela aplicação com o trecho correspondente no texto.

O valor de CF_{FMT} é calculado extraindo-se a média aritmética dos valores de seus modificadores. Desta forma, quando se trata do reconhecimento de referências do tipo *nome de lugar*, único caso em que o CF_{FMT} possui mais de um modificador (possui dois), o valor de CF_{FMT} corresponde à média dos valores obtidos para a taxa de abreviação e de utilização de iniciais maiúsculas. Para os demais casos, onde há apenas um único modificador, o CF_{FMT} assume o mesmo valor de CM_{FMT}.

4.2.6. Cálculo do Valor Final de Confiança

Obtidos os valores dos diversos fatores de confiança disponíveis, o valor final de confiança (CR) é calculado para cada referência detectada no processo de *parsing*. A partir deste, podem ser eliminadas as referências que não atingem o valor mínimo exigido, bem como selecionadas as referências mais confiáveis dentre as envolvidas em processos de eliminação de ambiguidade.

O valor de CR é obtido conforme exibido na Equação 4.5. Observa-se que o valor de CR representa o somatório do valor de cada CF associado à referência, ponderando-se de acordo com seu respectivo peso. Os CFs utilizados variam de acordo com o local de detecção da referência, com o tipo de referência detectada e com o tipo de lugar referenciado, de forma que para cada caso há determinados fatores característicos. Por exemplo, não faz sentido verificar o uso de maiúsculas em referências do tipo CEP. No ANEXO II relacionam-se os CFs utilizados em cada um dos casos possíveis, com seus respectivos pesos. Tais valores foram obtidos empiricamente, executando-se o processo de detecção de referências geográficas para uma coleção de dezenas de documentos contendo os diversos casos possíveis, avaliando-se então a qualidade dos resultados obtidos.

$$CR = \sum_{i=1}^{n} (CF_i \bullet P_i)$$ Equação 4.5

Onde:

- *CF_i é o o i-ésimo CF associado à referência;*
- *P_i é o peso de CF_i no valor final;*
- *n é o quantidade de CFs associados à referência analisada;*
- *$0 \leq CR \leq 1$.*

4.3. Modelagem do Escopo Geográfico

Após o processo de identificação das referências geográficas, tem início o processo de modelagem do escopo geográfico, cujo objetivo é mensurar o grau de relevância de cada localidade geográfica em relação ao documento. Os escopos dos documentos analisados podem conter tanto localidades referenciadas diretamente (identificadas explicitamente no documento através do processo anterior), quanto outras localidades derivadas a partir destas. Um documento pode ter escopo múltiplo, ou seja, um mesmo documento pode ser associado a mais de uma referência geográfica, cada uma dessas com seu respectivo grau de relevância ao documento.

O processo de modelagem mencionado explora a hierarquia geográfica *cidade → microrregião → mesorregião → estado → região*, de forma a gerar o escopo e calcular a relevância para níveis mais altos desta, a partir de referências encontradas em níveis inferiores. A este método foi atribuído o nome de e*xpansão do georreferenciamento*. Por exemplo, suponha que num documento foram identificadas referências a dois estados contidos em uma mesma região. Neste caso, além dos dois estados referenciados diretamente, poderá fazer parte do escopo geográfico deste documento a região que os contém, com seus respectivos graus de relevância. Para o cálculo dos valores de relevância, este método considera estatísticas sobre

as referências identificadas (e.g., número de ocorrências de uma referência no texto), estatísticas extraídas da hierarquia (e.g., quantidade de estados de uma região) e relacionamentos espaciais entre as localidades analisadas (e.g., dispersão geográfica).

Os objetivos principais da expansão do georreferenciamento são: transferir para o tempo de *parsing* e indexação algumas operações espaciais que, de outra maneira, seriam realizadas em tempo de realização de buscas; e proporcionar a atribuição da relevância geográfica de maneira mais completa, disseminando-a para outras localidades relacionadas e tornando possível obter, em tempo de execução de buscas, valores previamente calculados, mesmo para localidades indiretamente referenciadas. Maiores detalhes sobre a utilização dos resultados do processo de expansão são mostrados na seção 4.4, que explica o mecanismo de indexação espaço-textual. A seção a seguir apresenta a estrutura de dados elaborada para representação e manipulação das informações geradas pelo referido processo de expansão.

4.3.1. Geotree

A geotree é uma estrutura de dados em forma de árvore desenvolvida para a implementação da técnica de expansão do georreferenciamento, onde seus nós representam as localidades que fazem parte do escopo geográfico do documento analisado. A sua utilização proporciona maior flexibilidade ao processo de busca, fazendo com que qualquer localidade especificada na busca, mesmo que não tenha sido referenciada diretamente por um documento indexado, possa ser recuperada no escopo geográfico de um documento de forma mais direta, sem que seja necessário comparar espacialmente o argumento de busca especificado com o *footprint* espacial de cada documento indexado em tempo de execução.

Para ilustrar, suponha a existência de um documento *D* que possui referências a duas cidades *A* e *B* contidas espacialmente na microrregião *M*. Considere ainda uma busca onde foi especificado como argumento espacial a microrregião *M*, utilizando-se o operador espacial *inside*. Deste modo, como o documento *D* também é indexado pela referência à *M* (resultante do processo de expansão), e não somente pelas referências a *A* e *B*. O registro do índice relativo à *D* é retornado sem que seja necessário aplicar o operador espacial entre a geometria de *M* e o *footprint* espacial de cada registro do índice. O valor de relevância, por sua vez, é obtido imediatamente, visto que já havia sido previamente calculado para a localidade *M*

durante o processo de expansão. O mecanismo de realização de buscas está descrito na seção 4.6. A Figura 4.2 exibe uma instância de uma geotree, que será utilizada para auxiliar a descrição desta estrutura.

Uma geotree *T* é uma árvore (com nó raiz identificado por *root[T]*) que tem as propriedades a seguir:

1. Todo nó *x* tem os seguintes atributos:
   a. *location[x]*, uma localidade pertencente ao escopo geográfico do documento representado por *T*;
   b. *depth[x]*, a distância relativa do nó *x* ao nó *root[T]*;
   c. *type[x]*, o tipo do nó *x*, que pode ser *D*, *I* ou *H*, significando, respectivamente, que *location[x]* foi detectada diretamente, indiretamente ou de forma híbrida (direta e indiretamente) no processo de detecção de referências geográficas, integrante do processo de *parsing*;
   d. *weight[x]*, peso do nó *x*, que representa a importância, no documento, da localidade representada por esse nó, com base na quantidade de ocorrências dessas localidades dentre as referências detectadas diretamente e no relacionamento hierárquico entre o nó *x* e os demais nós;
   e. *weight-bal[x]*, peso ajustado de x, explicado mais adiante nesta seção;
   f. *dispersion-rate[x]*, medida de dispersão geográfica, descrita na seção 4.3.2;
   g. *geo-relevance[x]*, a relevância da localidade representada pelo nó *x* no documento, descrita na seção 4.3.3
2. Todo nó folha possui *type[x] = D*.
3. Todo nó não-folha, em que *location[x]* não tenha sido detectada diretamente, possui *type[x] = I*.
4. Todo nó não-folha, em que *location[x]* tenha sido detectada diretamente, possui *type[x] = H*.
5. O pai de um nó *x* é tal que *location[x]* esteja contida espacialmente em *location[parent[x]]*, exceto a raiz que não possui pai.
6. *O atributo depth[x]* representa inversamente o nível hierárquico de *location[x]*, ou seja, na hierarquia de cidade à região adotada nesta implementação, o nível mais

alto (região) representa o nó de menor *depth[x]*, o nó *root[T]*, enquanto o(s) nó(s) de maior *depth[x]* representa(m) o menor nível hierárquico dentre as localidades que fazem parte do contexto geográfico do documento.

7. Uma geotree possui altura (*height[T]*) máxima igual a *NUM_LEVEL-1*, sendo *NUM_LEVEL* a quantidade de níveis pertencentes à hierarquia geográfica. Desta forma, no caso da hierarquia utilizada no GeoSEn, as geotrees instanciadas possuem altura máxima igual a 4.

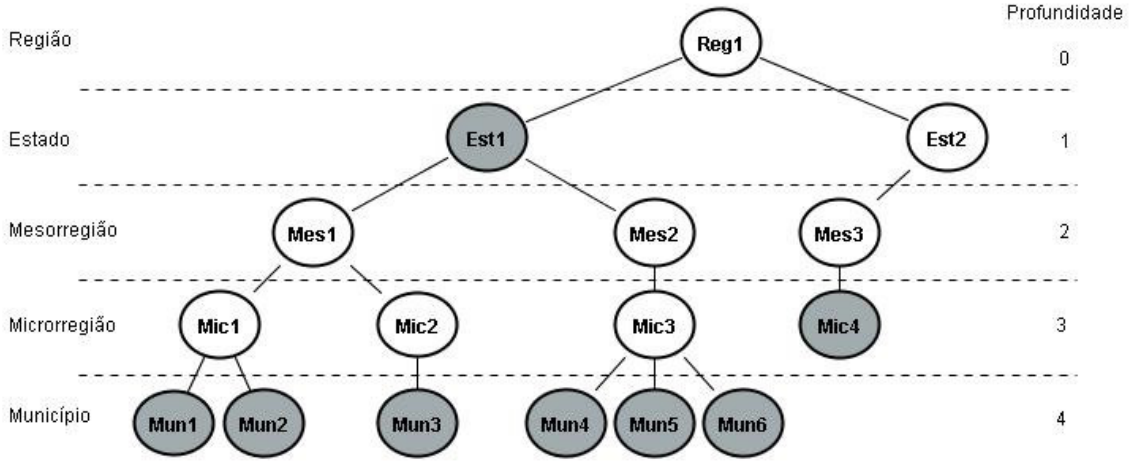


Figura 4.2 – Instância de uma Geotree

O peso de um nó é dado pela Equação 4.6, onde calcula-se o quociente entre o somatório dos pesos dos filhos do nó e o número de total de filhos que podem existir para aquele nó.

$$weight[x] = \begin{cases} \dfrac{\sum_{i=1}^{n[children[x]]} weight[child_i[x]]}{total-children\,[x]} + nRef & \text{, caso } x \text{ seja um nó não-folha} \\ nRef & \text{, caso } x \text{ seja um nó folha} \end{cases} \quad \text{Equação 4.6}$$

Onde:

- *$child_i[x]$ é o i-ésimo filho do nó x;*
- *n[children[x]] é a quantidade de filhos de x;*
- *total-children[x] é a quantidade máxima de filhos possíveis para x;*
- *nRef é a quantidade de referências R sendo L(R) = location[x]*

- *L(R) é a localização geográfica representada por uma referência R;*

Suponha dois documentos: D_1, contendo uma referência para cada município *A*, *B* e *C*; e D_2, contendo apenas uma referência ao município *A*. É de se esperar que o município *A* seja mais importante para D_2 do que para D_1, pois D_2 cita exclusivamente essa localidade, enquanto D_1 cita outras duas (*B* e *C*). Todavia, de acordo com a Equação 4.6, o nó representando o município A teria o mesmo peso em ambas as geotrees geradas para cada documento. Assim, com o objetivo de se obter valores mais coerentes para os pesos dos nós, os valores calculados segundo a Equação 4.6 são balanceados de acordo com a quantidade de nós do tipo *D* na mesma profundidade do nó. O peso balanceado de um nó *x* é identificado por *weight-bal[x]*. Este cálculo é exibido na Equação 4.7. Portanto, para fins do cálculo de relevância do documento (descrito mais adiante neste capítulo), é utilizado sempre o valor balanceado.

$$weight\text{-}bal[x] = \frac{weight[x]}{nY} \qquad \text{Equação 4.7}$$

Onde:

- *nY é o número de nós y, tal que* $type[y] \in \{D, H\}$ *e* $depth[y] = depth[x]$

Na Figura 4.2, os nós na cor cinza representam localidades detectadas diretamente, e os na cor branca as detectados indiretamente. Observam-se, na figura, sete nós do tipo *D* (*Mun1, Mun2, Mun3, Mun4, Mun5, Mun6 e Mic4*), classificados desta forma por suas referências terem sido detectadas diretamente e por serem folhas na árvore, ou seja, por não terem sido derivados de nenhum outro nó. Nota-se ainda a existência de apenas um nó do tipo *H* (*Est1*), por sua referência ter sido detectada diretamente e por não ser folha, visto que sua referência também foi derivada das referências dos nós *Mes1* e *Mes2*. Os demais nós são todos do tipo I, cujas referências foram apenas derivadas de outras, em nós do nível inferior.

Para representar todas as localidades referenciadas em um documento, utiliza-se uma ou mais geotrees. Como o nó raiz representa uma localidade pertencente ao nível mais alto da hierarquia, são instanciadas tantas geotrees quantas forem as localidades referenciadas que pertençam a este nível. Portanto, o número máximo de geotrees instanciadas por documento é

igual ao número de localidades existentes no maior nível da hierarquia adotada. Com a hierarquia definida neste trabalho (de cidade à região), o número máximo de geotrees por documento é igual cinco, dado a existência de cinco regiões no Brasil. No caso da utilização de uma hierarquia em escala global, por exemplo, este número poderia ser seis, representando os seis continentes existentes no planeta. O Código 4.1 apresenta o algoritmo *GEOTREE*, que descreve em pseudocódigo os passos necessários para a construção das geotrees relacionadas a um documento.

```
GEOTREE(geoscope)
1       for (i ← 1 to geoscope.locations.length)
2               location ← geoscope.locations[i]
3               depth ← NUM_LEVEL - location.level
4               type ← D
5               weight ← recuperarQuantidadeReferencias(geoscope, location)
6               node ←  criarNo(location, depth, type, weight)
7               adicionarNo(geotree, node)
8
9       maxDepth := NUM_LEVEL-1
10      for (i ← maxDepth to 1)
11              nodesLevel ← recuperarNosPorNivel(geotree, i)
12              for (j ← 1 to nodesLevel.length)
13                      node ← nodesLevel[j]
14                      parentLocation ← recuperarLocalidadeSuperior(node)
15                      parentNode ← recuperalNoPelaLocalizacao(geotree, parentLocation)
16                      if (parentNode = null)
17                              maxFilhos ← recuperarMaximoFilhos(parentLocation)
18                              parentNode ← criarNo(parentLocation, i-1, I, node.weight/maxFilhos)
19                              adicionarNoFilho(parentNode, node)
20                              adicionarNo(geotree, parentNode)
21                      else
22                              adicionarNoFilho(parentNode, node)
23                              atualizarNo(geotree, parentNode)
24
```

```
25      atualizarGeoScope(geoscope, geotree)
26      return geotree
```

Código 4.1 – Algoritmo de construção de uma geotree

As referências detectadas diretamente são inseridas em uma estrutura denominada *geoscope*, que representa o escopo geográfico do documento. Nesta, as referências são agrupadas por suas localidades, ou seja, cada grupo contém um conjunto de referências para uma mesma localidade. Então, o algoritmo apresentado recebe esta estrutura como parâmetro e itera sobre estas localizações (linha 1), criando um nó para cada localidade distinta (linhas 2 a 6) e em seguida adicionando o novo nó à geotree (linha 7). Tais nós inicialmente criados possuem os demais atributos: tipo D; profundidade igual a *NUM_LEVEL - location.level*, onde *location.level* é o nível hierárquico da localidade analisada; e o peso é dado pela quantidade de referências detectadas para aquela localidade, conforme descrito na Equação 4.6.

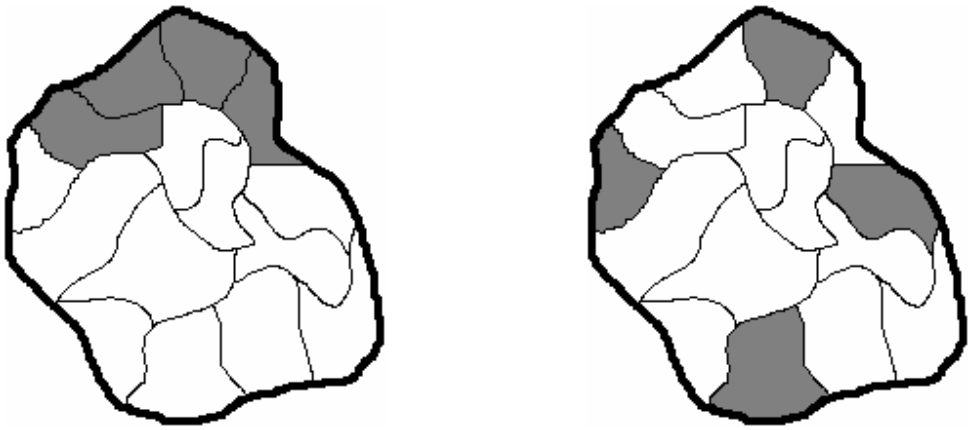
Em uma segunda etapa, o algoritmo itera sobre a *geotree* inicialmente gerada, do nível mais profundo até o nível 1, onde estão localizados os nós imediatamente abaixo das raízes, ou seja, *children[root[T]]* (linha 10). Então, o algoritmo itera sobre os nós de cada nível (linha 12). Para um determinado nó, executa-se uma consulta espacial para recuperar a localidade cujo nível hierárquico seja imediatamente superior ao do nó e que contenha espacialmente a localidade por este representada (linha 14). A partir da localidade retornada, recupera-se o nó da *geotree* que a representa. Este nó deve ser o ancestral direto do nó analisado. Se este nó ancestral ainda não existir (linhas 17 a 20), ele é criado e adicionado à geotree. Neste caso, o nó criado tem como atributos: a localidade recuperada na linha 14; a profundidade igual a i-1, onde i é a profundidade do nó analisado; o tipo *I*; e o peso inicial igual 1/(total de filhos possíveis). Por outro lado, caso o nó ancestral já exista (linhas 22 a 23), o nó analisado é adicionado aos filhos do nó ancestral e em seguida este é atualizado na geotree, com um novo valor calculado para seu peso, relativo à adição do novo filho (vide equação Equação 4.6). Por fim, atualiza-se o *geoscope* com as novas referências resultantes da execução do algoritmo.

4.3.2. Dispersão Geográfica

A medição da dispersão geográfica, outro método desenvolvido neste trabalho, visa mensurar o quão dispersas estão as localidades geográficas integrantes do escopo geográfico de um documento. O valor resultante é um dos fatores determinantes na formação do valor da relevância geográfica calculado para cada localidade pertencente ao escopo.

O grau de dispersão é calculado para todo nó *x* de uma *geotree* (atributo *dispersion-rate[x]*), representando a dispersão geográfica, dentro dos limites da localidade representada por *x*, das localidades referenciadas por seus nós descendentes. O valor de *dispersion-rate[x]* está compreendido no intervalo de 0 a 1. Deste modo, este método considera mais relevante uma localidade cujos nós filhos estejam mais espalhados geograficamente. Desta forma, estabeleceu-se que os nós do tipo *D* e *H* possuem *dispersion-rate[x]* igual a 1, o valor de dispersão máxima.

Para ilustrar, suponha um documento onde são referenciados quatro estados contidos em uma mesma região *R*. A geotree resultante conterá cinco nós, sendo quatro do tipo *D*, representando cada um dos estados referenciados diretamente, e um do tipo *I*, representado a região que contém tais estados. Duas possibilidades para este exemplo são exibidas na Figura 4.3, onde os estados detectados estão destacados em tonalidade mais escura. Na Figura 4.3a, os quatro estados referenciados estão concentrados ao norte da região *R*. Já na Figura 4.3b, os quatro estados estão distribuídos por toda a região. Desta forma, o método da dispersão considera mais relevante para a região *R* o documento da Figura 4.3b.



(a) (b)

Figura 4.3 – Exemplos de dispersões geográficas.

Esta abordagem tem como hipótese que referências pouco espalhadas (pouco dispersas geograficamente), em geral, caracterizam sub-regiões (oficiais ou não), que compartilham características comuns como, por exemplo, culturais, econômicas, sociais, climáticas, dentre outras, com menor probabilidade de representarem a região como um todo. Por outro lado, referências mais espalhadas em uma mesma região, denotam uma representação mais abrangente, com maior probabilidade de estarem relacionadas à região de forma geral. Para exemplificar, admita dois documentos em que ambos discutam sobre futebol e possuam referências a cinco cidades brasileiras. Porém, no primeiro documento, todas as cidades estão localizadas no sertão pernambucano; já no segundo, as cidades estão distribuídas entre as regiões sul, sudeste e nordeste do Brasil. Assim, é de se esperar que o segundo documento possua um contexto mais nacional que o primeiro, que referencia uma região mais restrita do país. Nesta situação, o primeiro documento poderia ser, por exemplo, uma reportagem sobre um projeto de incentivo ao esporte, promovido pelas prefeituras locais em parceria com uma instituição filantrópica de atuação regional. Enquanto o segundo poderia ser, por exemplo, uma reportagem sobre a última rodada do campeonato brasileiro.

Foi visto que se um nó *X* é do tipo *D* ou *H*, seu atributo *dispersion-rate[x]* é igual a 1. No entanto, caso seja do tipo *I*, este valor é calculado da seguinte forma: recupera-se um conjunto *C* contendo os nós do tipo *D* descendentes de *X*. Obtém-se o *envelope* E_1, que cobre todas as localidades associadas aos nós de *C*, e o *envelope* E_2, que cobre a localidade referenciada por *X*. Obtidos os valores de E_1 e E_2, o valor final da dispersão para *X* pode ser obtido aplicando-se a Equação 4.8. A Figura 4.4 ilustra a aplicação dos *envelopes* E_1 e E_2, destacados em retângulos tracejados.

$$dispersion\text{-}rate[x] = \frac{Área(E1)}{Área(E2)} \qquad \text{Equação 4.8}$$

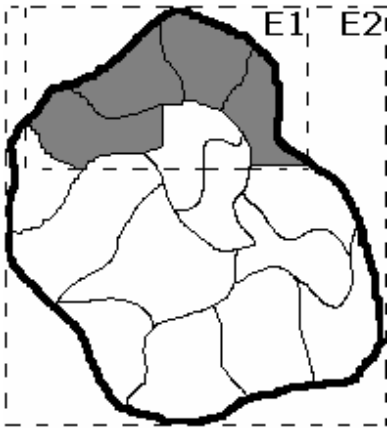


Figura 4.4 - Exemplo de aplicação dos envelopes para o cálculo da dispersão geográfica.

4.3.3. Relevância Final

A relevância geográfica é calculada para cada nó *X* de uma *geotree* (geo-*relevance[x]*), a partir dos valores de *weight-bal[x]* e *dispersion-rate[x]*, segundo a Equação 4.9.

$$\text{geo-}relevance[x] = weight\text{-}bal[x] \bullet (1 + dispersion\text{-}rate[x]) \qquad \text{Equação 4.9}$$

Como pode ser observado no cálculo de geo-*relevance[x]*, o valor *dispersion-rate[x]* é utilizado apenas para ajustar positivamente o valor de *weight-bal[x]*, ou seja, por menor que seja o valor de *dispersion-rate[x]*, este nunca influenciará negativamente o cálculo do valor final de relevância. Desta forma, se *dispersion-rate[x]* = 0, implica geo-*relevance[x]* = *weight-bal[x]*.

4.4. Indexação Espaço-textual

Como foi visto em capítulos anteriores, os sistemas de busca geográfica necessitam indexar os documentos recuperados por suas perspectivas textual e espacial. No âmbito textual, a maioria das soluções utilizadas no mercado baseia-se em arquivo invertido. No GeoSEn não é diferente, visto que este indexa os termos de seus documentos utilizando o

índice no formato do Apache Lucene. No domínio espacial, ainda existe certa variação nas soluções apontadas na literatura. Em sua maioria, essas soluções associam cada documento indexado a um *footprint* espacial, que é a representação espacial das localidades por este referenciadas. As soluções de indexação espacial se diferenciam, principalmente, pela indexação de um escopo simples ou múltiplo (respectivamente, significam que apenas uma única localidade é associada a um documento ou que várias localidades podem ser associadas a um mesmo documento), e pela forma de representação do *footprint* espacial (e.g., uma geometria do tipo polígono ou apenas um ponto representando o centróide).

A abordagem adotada neste trabalho para o processo de indexação espacial utiliza como base o resultado do processo de expansão do georreferenciamento. Assim, cada localidade referenciada direta ou indiretamente pelo documento possui uma entrada independente no índice, associada a sua respectiva relevância geográfica, ou seja, existe uma entrada no índice para cada nó da *geotree* gerada para o documento. Como foi visto anteriormente, o índice do Lucene oferece possibilidade de extensão, ou seja, de adição de novos metadados associados a cada documento indexado. Desta forma, conforme mencionado na seção 4.1, sobre a arquitetura do sistema, foi desenvolvido um *plugin* chamado *GeoSEn Indexing Filter*, responsável por adicionar ao índice informações espaciais acerca dos documentos. Este *plugin* cria, no registro indexado do documento, um novo campo denominado *geoscope*, que contém os códigos de identificação (ID) das localidades referenciadas pelo documento, com seus respectivos graus de relevâncias.

É certo que este método ocasiona um maior crescimento da quantidade de registros contidos no índice, uma vez que cada documento não está mais associado apenas às localidades referenciadas diretamente, mas também a outras que foram detectadas indiretamente através do método de expansão. Entretanto, como os documentos são indexados através dos IDs das localidades ao invés dos *footprints* espaciais, o índice torna-se menor em relação ao tamanho ocupado em disco, apesar da maior quantidade de registros.

Com este modelo, é possível reusar grande parte do mecanismo de processamento de consultas do Nutch para consultar os dados espaciais associados aos documentos indexados, de forma similar ao que acontece com os parâmetros textuais de busca. O método desenvolvido para o processamento de consultas nesta estrutura de indexação está descrito na seção 4.6.

4.5. Interface Multi-modo

A Figura 4.5 exibe a tela principal da interface gráfica construída para o protótipo. Nessa interface, estão disponíveis aos usuários os modos de interação textual e espacial. O modo textual é similar aos encontrados em outros sistemas de busca disponíveis no mercado, com a entrada de dados realizada através de alguns campos de formulário e o resultado da busca exibido através de uma lista de documentos recuperados, contendo alguns dados acerca dos documentos, como o título, a URL e um fragmento de texto. No modo espacial, por sua vez, utiliza-se um mapa interativo, onde é possível realizar operações básicas de visualização como *pan*, *zoom* e controle de camadas, além da seleção visual das localidades a serem utilizadas na busca e da visualização de seus resultados. É possível selecionar localidades pré-fixadas (e.g., uma mesorregião) ou delimitar regiões retangulares livremente. Para implementação do mapa, foi utilizada a API Google Maps versão 2.0 (code.google.com/apis/maps/).

Figura 4.5 – Tela principal do GeoSEn

No campo “argumento textual”, apontado na Figura 4.5, informam-se os termos pelos quais serão recuperados os documentos. No campo “operação espacial”, é selecionada a operação espacial desejada. Atualmente, estão disponíveis as operações de continência (i.e., lugares contidos nas regiões especificadas), de distância (i.e., lugares a uma certa distância das regiões especificadas), de adjacência (i.e., lugares adjacentes a uma localidade especificada) e suas respectivas negações. Estas operações são referenciadas neste documento, respectivamente, por *inside*, *distance* e *adjacency*; e suas negações são precedidas de “*not* +”. Os operadores de negação são selecionados ativando-se o campo “não”. O campo “distância” torna-se disponível quando é selecionada a operação *distance*. Neste, deve ser informado o valor de distância em quilômetros. No campo “localidade”, devem ser informados os nomes

(ou parte deles) das localidades a serem utilizadas no processo de busca. Podem ser especificadas várias localidades, separando-as por ponto-e-vírgula. Ainda na Figura 4.5, pode-se observar no mapa uma situação de exemplo, onde foi selecionado o estado da Paraíba, e onde o cursor do mouse encontra-se posicionado sobre o estado de Alagoas, cujo nome é exibido na parte inferior do mapa. Neste caso, o sistema destacou ambos os estados no mapa, utilizando cores distintas.

A Figura 4.6 mostra a tela de eliminação de ambiguidade do argumento informado no campo localidade da tela principal, exibida na Figura 4.5. Na tela de eliminação de ambiguidade, são disponibilizados para seleção pelo usuário os lugares disponíveis para cada nome especificado.

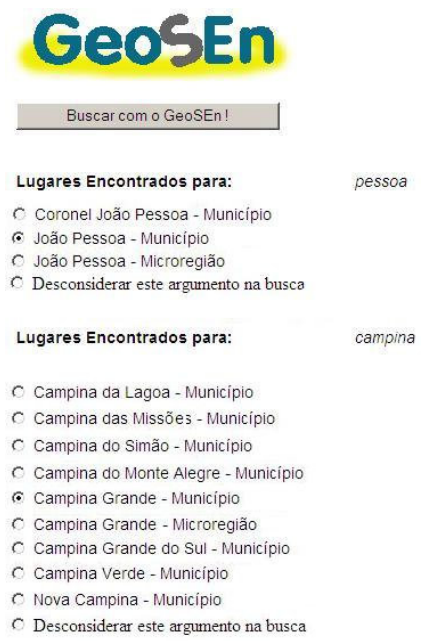


Figura 4.6 – Tela eliminação de ambiguidade das localidades informadas para busca

Por fim, exibe-se na Figura 4.7 a tela onde são mostrados os resultados das buscas. Para cada item do resultado, há ainda três links – *Cache*, *Explain* e *Geo Explain* – utilizados para visualizar, respectivamente: uma cópia do conteúdo do documento mantida na base de dados; maiores detalhes sobre a geração do *ranking* de relevância textual (de acordo com os argumentos de busca especificados); e detalhes sobre o escopo geográfico do documento listado. Os recursos de *Cache* e *Explain* são disponibilizados pelo Nutch, desenvolvidos para

prover maior transparência ao usuário sobre o processo utilizado na seleção dos documentos. Seguindo este princípio, foi desenvolvido o recurso *Geo Explain* para o GeoSEn, promovendo aos usuários maior clareza sobre o processo de georreferenciamento dos documentos consultados.

Figura 4.7 – Tela de exibição dos resultados de busca

4.6. Execução de Buscas

Foi visto na seção 4.1, que trata da arquitetura do sistema, que para o reconhecimento de novos campos adicionados ao índice do Lucene, faz-se necessária a implementação de plugins específicos do Nutch. Com este fim, foi desenvolvido o plugin *Geosen Query Filter*, que é capaz de selecionar os documentos do índice que contêm determinados valores no campo *geoscope*, descrito na seção 4.4. Desta forma, conforme explicado posteriormente nesta

seção, a dimensão espacial das buscas solicitadas pelos usuários é convertida em um conjunto de IDs de localidades. Estes IDs, por sua vez, são utilizados pelo referido *plugin* para filtrar os documentos que estão ou não associados a tais localidades, de forma análoga aos *plugins* responsáveis por filtrar os documentos que estão associados aos termos especificados na dimensão textual da consulta. Portanto, observa-se que, uma vez convertida a representação espacial da busca em um conjunto de IDs, o restante do processo é reusado do Nutch, fazendo com que o sistema desenvolvido desfrute da escalabilidade e eficiência no processamento de consultas asseguradas pelo Nutch.

Através da seleção de localidades pré-fixadas, a recuperação dos IDs correspondentes é direta. No entanto, quando a região é especificada através de seleção livre (e.g., seleção retangular), ou quando são utilizados os operadores *distance* e *adjacency*, faz-se necessária a execução de um procedimento de análise espacial para relacionar os IDs correspondentes. Dada uma seleção retangular a ser aplicada ao operador *inside*, são selecionadas as localidades de maior nível hierárquico possível, dentre as contidas na área especificada, eliminando-se as redundâncias (localidades que estão contidas umas nas outras).Um esboço simplificado da consulta SQL utilizada neste procedimento é exibida no Código 4.2. Nesta, são recuperados os IDs das localidades (linha 1) cujas geometrias estejam contidas na região especificada (linha 4) e que não estejam contidas na geometria de nenhuma outra localidade também contida nesta região (linhas 5 a 10).

```
1   SELECT id
2   FROM places plc1
3   WHERE
4     within(plc1.geometry, specified_geometry)
5     AND NOT EXISTS (
6           SELECT id
7           FROM places plc2
8           WHERE
9             within(plc2.geometry, specified_geometry)
10            AND within(plc1.geometry, plc2.geometry)
11    )
```

Código 4.2 – Consulta para recuperar os IDs a partir de uma geometria retangular

No caso de busca por distância com seleções retangulares, o procedimento é análogo; porém, a região especificada no início do processo é ampliada utilizando-se a função espacial *buffer*, tendo como parâmetro o valor de distância especificado pelo usuário. Em seguida, a nova geometria retangular é submetida à consulta do Código 4.2. Já para o caso da aplicação do operador *distance* com seleção de localidade pré-fixada, calcula-se inicialmente o MBR desta e, de posse de uma geometria retangular, aplica-se o processo descrito para o caso de seleções livres retangulares.

A Figura 4.8 ilustra as etapas envolvidas no processo descrito. Nesta, exibe-se o mapa de uma região fictícia, contendo três níveis hierárquicos. Considere que estes níveis sejam de cidade, estado e região. Para facilitar a visualização, foram utilizadas linhas de diferentes espessuras para representar os limites entre as localidades de diferentes níveis hierárquicos. Os limites do nível de cidade são representados em linhas finas, os de estado em linhas intermediárias e os de região em linhas grossas. Em (a), mostra-se o mapa de exemplo com suas respectivas localidades delimitadas. Em (b), exibem-se uma suposta região retangular definida para a execução de uma busca utilizando o operador *inside* e as localidades selecionadas internamente pelo sistema. Neste exemplo, foram selecionadas 10 localidades dentre as 67 que intersectam o retângulo, sendo 1 região, 2 estados e 7 cidades. Foram eliminadas as localidades que não estão completamente contidas no retângulo e as localidades mais internas (de nível hierárquico inferior) em relação a alguma outra localidade selecionada. A eliminação das localidades de níveis inferiores acontece pelo fato destas serem dadas como redundantes, uma vez que pelo método de expansão do georreferenciamento, todo documento associado a uma localidade *x*, qualquer localidade *y* contendo espacialmente *x* é adicionada também ao escopo geográfico do documento.

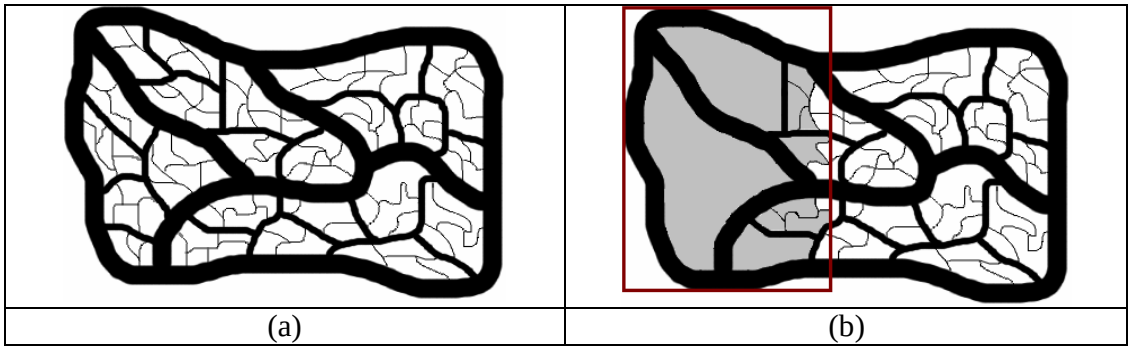


Figura 4.8 - Seleção de IDs a partir de uma região retangular especificada

A utilização do operador *adjacency* só é permitida quando são especificadas localidades pré-fixadas, ou seja, não é possível utilizar este operador com a seleção de regiões retangulares. Dada uma busca utilizando-se este operador, selecionam-se os IDs das localidades adjacentes à localidade especificada pelo usuário e que sejam do mesmo nível hierárquico desta. Para ilustrar a utilização deste procedimento, considere o exemplo da Figura 4.9. Neste, há dois estados E1 e E2, adjacentes, contendo 7 e 6 cidades, respectivamente. Suponha que D seja um documento com referência apenas à cidade C11. De acordo com o processo de expansão do georreferenciamento, o estado E1 estará contido no escopo geográfico de D. Assim, dada uma busca onde foram especificados o estado E2 e o operador *adjacency*, o documento D seria recuperado. Por outro lado, caso tenha sido especificada na busca apenas a cidade C11, um documento contendo referência a E2 não seria necessariamente recuperado, mesmo sendo C1 e E2 adjacentes, uma vez que não são de mesmo nível hierárquico. Neste caso, seriam recuperados os documentos com referência às cidades C12, C15, C21, C22 ou C23.

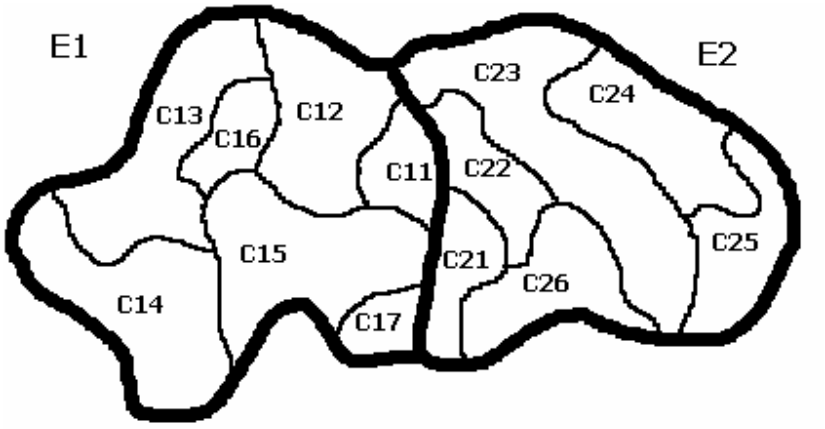


Figura 4.9 - Exemplo de utilização do operador *adjacency*

Em geral, quando os documentos são indexados pelo seu *footprint* espacial, não é necessário nenhum procedimento de consulta espacial antes de submeter a consulta ao índice, entretanto, as análises espaciais são realizadas comparando-se a região especificada na busca com o *footprint* espacial de cada registro do índice, ocasionando maior custo em relação ao procedimento apresentado, uma vez que a quantidade de registros no índice é muito maior que a quantidade de localidades cadastradas no sistema. Observe que o método de indexação e

consulta utilizando IDs torna-se possível devido ao processo de expansão do georreferenciamento, onde são pré-calculadas, durante a execução do parsing, as relevâncias geográficas até o nível mais alto da hierarquia, em cada documento, fazendo com que não seja necessário executar consultas espaciais a cada registro do índice em tempo de realização de busca.

Para cada item resultante da execução da busca, é calculado um valor de relevância em relação aos argumentos especificados. Esta medida é composta pelos valores de relevância textual e espacial do documento em relação à busca, atribuindo-se um peso específico a cada um deles. O peso atualmente atribuído para ambos é de 50%, ou seja, confere-se igual importância a cada uma das perspectivas de busca. Todavia, estes valores podem ser facilmente alterados pelo administrador do sistema. O grau de relevância textual é o mesmo calculado pelo Nutch. Já o grau de relevância espacial é obtido através da soma dos valores de relevância previamente calculados para as localidades associadas ao documento e que também façam parte da coleção de IDs gerada para consultar o índice.

Note que, quanto mais localidades associadas um documento possuir dentre as especificadas como argumento de busca, maior a quantidade de elementos somados e, consequentemente, maior a probabilidade de se obter um valor maior para o resultado da soma. Em seguida, os valores resultantes da referida soma são normalizados para então serem utilizados no cálculo do valor final de relevância. Por fim, os itens do resultado são ordenados segundo seu grau de relevância final e então exibidos ao usuário.

Para ilustrar o processo descrito, suponha um índice formado conforme a Tabela 4.2. O suposto índice contém 4 documentos, cujos códigos de identificação são exibidos na coluna Document ID. Cada um desses documentos estão associados a 4 localidades, para as quais são exibidos seus IDs e seus valores de relevância em relação ao documento, separados por “/”.

Document ID	**Place ID / Relevance**			
Doc001	001 / 1.256	003 / 1.654	005 / 2.343	007 / 3.445
Doc002	002 / 4.322	004 / 1.881	006 / 1.934	008 / 1.122
Doc003	001 / 2.562	002 / 3.210	003 / 1.420	004 / 1.693
Doc004	002 / 3.210	003 / 1.420	004 / 1.872	005 / 2.333

Tabela 4.2 - Exemplo de conteúdo do índice

Desse modo, considere uma busca submetida ao suposto índice, contendo como argumento espacial os IDs 001, 002 e 003. Para esta, recuperam-se os documentos cujos escopos contêm algum dos IDs especificados, somando-se os respectivos valores de relevância. Neste exemplo de busca, todos os documentos do índice seriam retornados. A Tabela 4.3 exibe, respectivamente, os IDs dos documentos retornados, os IDs das localidades especificadas na consulta que foram encontrados em cada documento, os valores de relevância utilizados na composição do valor final, o valor total calculado e o valor de relevância normalizado.

Document ID	IDs encontrados no índice	Valores utilizados no somatório	Resultado do somatório	Resultado Normalizado
Doc001	001, 003	1.256 + 1.654	2.910	0.405
Doc002	002	1.322	4.322	0.600
Doc003	001, 002, 003	2.562 + 3.210 + 1.420	7.192	1.000
Doc004	002, 003	3.210 + 1.420	4.630	0.644

Tabela 4.3 - Exemplo de cálculo de relevância reográfica

Observe que os valores de relevância para as localidades 002 e 003 são iguais nos documentos Doc003 e Doc004; entretanto, para o exemplo de busca dado, o documento Doc003 obteve um valor mais alto de relevância, uma vez que está associado a mais lugares especificados na consulta do que o Doc004. Por outro lado, observa-se que o Doc002 contém menos lugares coincidentes com a consulta do que o Doc001; todavia, a localidade associada ao Doc002 possui um alto valor de relevância, fazendo com que o valor final de relevância geográfica calculada para este documento seja maior do que o calculado para o Doc001.

4.7. Busca por Zonas Temáticas

Além das localidades que fazem parte da hierarquia de cidade à região, é possível realizar buscas utilizando como argumento espacial algumas localidades formadas por fatores sociais, econômicos, culturais e climáticos, como por exemplo, a busca “empresas de blindagem em regiões brasileiras com baixo índice de violência”. Neste exemplo, o argumento

espacial seria “regiões brasileiras com baixo índice de violência”. Estas são chamadas de *zonas temáticas*.

Para isto, são previamente carregados no sistema os devidos mapas temáticos, contendo seu nome e suas informações espaciais. Internamente, a região delineada no mapa é convertida em um conjunto de referências a localidades contidas na hierarquia (conjuntos de IDs), da mesma forma como são convertidas as regiões especificadas pelos usuários na processo de execução de buscas, descrito na seção 4.6. Estes mapas são devidamente anotados com algumas palavras-chave relacionadas ao tema. Desta forma, os termos utilizados pelos usuários para informar a região de interesse (no campo localidade da Figura 4.5) são comparados às palavras-chave dos mapas; então, da mesma forma como acontece com os nomes de localidades, são exibidos na tela de eliminação de ambiguidade (Figura 4.6) os nomes das zonas temáticas que satisfazem à consulta, para que o usuário selecione as desejadas e conclua o processo de busca.

4.8. Análise Comparativa

As principais contribuições deste trabalho concentram-se nos processos de modelagem e representação do escopo geográfico dos documentos, bem como no método de geração do ranking de relevância geográfico. No que se refere a estes mecanismos, podem-se dividir as pesquisas em GIR em duas linhas. A primeira, a exemplo dos trabalhos de Martins et al [35][38][48] e Jones et al [46][47][50], visa representar o escopo geográfico dos documentos através de um *footprint* espacial, indexando-os a partir de uma geometria que pode ser, por exemplo, um ponto ou um polígono. A outra linha, para a qual citam-se as pesquisas de Yi Li et al [36] e Amitay et al [41], concebe o escopo geográfico como um conjunto de localidades, que podem ser representas, por exemplo, por um conjunto de códigos identificadores. Para cada uma destas abordagens, relacionam-se aspectos positivos e negativos. Na primeira, há possibilidade de se realizarem buscas espaciais mais complexas, através do confronto entre os *footprints* da consulta e dos documentos indexados. Na segunda, é possível calcular (previamente à realização das buscas) valores de relevância para cada localidade associada a um documento, reduzir o custo de processamento de consulta.

De acordo com esta classificação, podem-se considerar as propostas deste trabalho como pertencentes à segunda linha. Os trabalhos de Yi Li et al [36], Zhisheng Li et al [42] e Amitay et al [41] se assemelham a este pela utilização de locais implícitos que, neste trabalho, são frutos do processo de expansão do georreferenciamento. No entanto, Yi Li et al [36] e Zhisheng Li et al [42] relatam dificuldade na realização de consultas utilizando operadores espaciais diferentes de "inside" em índices expandidos. Nestes casos, Yi Li et al [36] opta pela expansão de consultas; e Zhisheng Li et al [42] por uma estrutura de indexação baseada em um grid espacial. Já no presente trabalho, apresentam-se processos alternativos de buscas espaciais mais complexas utilizando a estrutura de índice expandido. Yi Li et al [36] relata ainda dificuldade em atribuir valores de relevância às localidades expandidas do índice. Um processo de atribuição destes valores de relevância é proposto por Amitay et al [41], bem como no presente trabalho. Contudo, tais propostas se diferenciam significativamente na forma como os referidos valores são calculados.

No Capítulo 3, foram apresentadas as importantes pesquisas em GIR, selecionados com base na quantidade de informações disponíveis na literatura. Através da Tabela 4.4 é possível comparar com maior facilidade as funcionalidades oferecidas pelo GeoSEn e as que são disponibilizadas nos protótipos desenvolvidos nestas pesquisas. Esta tabela difere da tabela Tabela 3.1 pela existência de uma coluna adicional referente às informações relacionadas ao GeoSEn. Da mesma forma como descrito para a tabela Tabela 3.1, a marcação ● denota a existência de uma característica em um projeto; e a ausência desta marca significa que a característica não está presente ou que esta informação não foi encontrada na literatura.

Descrição	**Geo Search**	**SPIRIT**	**GSE for Germany**	**Geo Tumba**	**GeoSEn**
Distingue entre escopo geográfico de uma página Web e locais de interesse por esta página.			●		
Considera fenômenos geográficos como lugares (e.g., um rio, uma montanha).		●			
Calcula valores de confiança para as referências geográficas detectadas no texto.			●	●	●
Considera alguns termos como especiais no processo de detecção de referências geográficas.	●	●	●	●	●

Descrição	Geo Search	SPIRIT	GSE for Germany	Geo Tumba	GeoSEn
Analisa a formatação dos elementos em documentos HTML (e.g., uso de negrito, itálico, coloração e outras *tags* de destaque).				●	
Analisa a grafia das referências no processo de detecção (e.g., uso de maiúsculas e abreviações nos nomes de lugares, forma como um cep está descrito, etc).					●
Possui uma base de dados com valores pré-calculados de probabilidade de um determinado nome referir-se a algum lugar por este identificado.					●
Para uma referência detectada, altera seu valor de confiança com base em relacionamentos espaciais com outras localidades diferentes também detectadas no texto.					●
Utiliza informações contidas nas URLs.	●		●	●	●
Utiliza informações do Whois.	●		●	●	
Utiliza estatística entre os termos e os textos (e.g. número de ocorrências, posicionamento no texto).	●			●	●
Permite associar escopos geográficos múltiplos aos documentos.		●			●
Utiliza estatísticas sociais (e.g., contingente populacional) nos processos relacionados à atribuição de escopo geográfico e elaboração do ranking de relevância.		●	●	●	
Utiliza ontologias para apoiar o processo de modelagem e manipulação de informações geográficas semânticas.		●		●	

Descrição	Geo Search	SPIRIT	GSE for Germany	Geo Tumba	GeoSEn
Analisa a estrutura de ligação entre as páginas da web (*links*).	●	●	●	●	
Atribui escopo geográfico e elaboração do ranking de relevância com base no padrão de distribuição dos locais referenciados.					●
Possui interface multi-modo.	●	●		●	●
Permite ao usuário balancear, em tempo de execução, entre a relevância geográfica e textual para elaboração do *ranking*.			●		
Utiliza técnicas de expansão de índice.					●
Calcula valores de relevância para as localidades expandidas do índice.					●
Disponibiliza três ou mais operadores espaciais para execução da busca.		●		●	●
Permite selecionar visualmente, como argumento espacial de busca, localidades pré-definidas e/ou especificar regiões por seleções livres.		●		●	●
Considera zonas temáticas como localidades válidas no processo de busca (e.g., região com maior índice de violência, região com maior renda per capita).					●
Suporta recuperação de recursos multimídia (e.g., imagens, vídeos) com base em seus escopos geográficos.					
Possui versão para aplicativos móveis.				●	

Tabela 4.4 – Comparativo entre as características verificadas no GeoSEn e em outros projetos de GIR.

4.9. Conclusão

Neste capítulo, foi possível conhecer o protótipo desenvolvido para validar as pesquisas realizadas neste trabalho. Foram apresentados os aspectos arquiteturais do sistema e o funcionamento dos seus principais componentes, realizando-se a correlação desses com as idéias pesquisadas e os métodos e técnicas elaboradas. O capítulo a seguir relata a metodologia utilizada e os resultados obtidos nas avaliações experimentais realizadas com o protótipo apresentado.

Capítulo 5
Avaliação Experimental

Neste capítulo, serão apresentados os resultados de alguns experimentos realizados com o protótipo descrito no Capítulo 4. Dentre os testes realizados, foram selecionados os que têm como objetivo mensurar a eficácia dos mecanismos de detecção de referências geográficas, de modelagem do escopo geográfico e de execução de buscas.

5.1. Procedimento Experimental

Para execução dos casos de teste relacionados à detecção e modelagem geográfica, foram selecionados manualmente diversos documentos da Web. A seleção de tais documentos baseou-se em seus respectivos conteúdos, os quais contêm situações variadas a serem avaliadas nos respectivos casos de teste. Para validação do mecanismo de execução de busca, foram indexados 66.531 documentos utilizando o robô contendo as funcionalidades implementadas para o protótipo.

Diante da necessidade de se limitar a quantidade de documentos a serem indexados para realização dos testes funcionais do protótipo, foram definidos alguns critérios para coleta destes documentos, de forma a se obter uma amostra da Web que fosse limitada a um conjunto restrito de categorias de conteúdo (visando dispor de maior controle sobre os testes realizados), porém preservada a aleatoriedade no que se diz respeito à contextualização geográfica de tal conteúdo.

Para cada uma das classes de conteúdo definidas, foi especificado um conjunto de URLs a serem utilizadas como raízes do processo de *crawling*, sendo estas independentes ao

máximo de quaisquer características geográficas. As classes definidas foram: notícias (e.g., reportagens em jornais e revistas), turismo (e.g., roteiros de viagem, hotéis, restaurantes, etc), governos (e.g., órgãos de governos municipais, estaduais e federais), universidades (e.g., instituições de ensino públicas e privadas), arte (e.g., música e dramaturgia) e blogs diversos.

5.2. Detecção de Referências Geográficas

Como já foi discutido nos capítulos anteriores, o grande desafio do detector de referências geográficas é resolver os diversos tipos de ambiguidades relacionadas ao processo. Os principais casos de ambiguidade estão sumarizados a seguir:

A. *Lugar com nome de coisa* (e.g., municípios Arame-Ma e Bugre-Mg);

B. *Lugar com nome de pessoa* (e.g., municípios Iracema-Ce e Cláudia-Mt);

C. *Lugares com mesmos nomes e mesmos tipos* (e.g., municípios Cachoeirinha-Pe e Cachoeirinha-Rs);

D. *Lugares com mesmos nomes e com tipos diferentes* (e.g., município Rio de Janeiro e estado Rio de Janeiro);

E. *Lugares com mesmos nomes de outros não cadastrados* (e.g., município pernambucano Buenos Aires, cadastrado, e o município argentino Buenos Aires, não cadastrado);

F. *Lugares e gentílicos com nomes iguais* (e.g., município Paulista-Pe e gentílico "paulista", relacionado ao estado de São Paulo);

Nos experimentos realizados com esse mecanismo, utilizou-se como referência os casos de ambiguidade listados, sendo avaliada a eficácia dos resultados obtidos com base no nível de resolução de cada um desses casos, que serão referenciados no texto através dos marcadores (letras) que foram utilizados para listá-los. Os documentos selecionados para realização dos experimentos relacionados a esse módulo do protótipo foram classificados de acordo com o nível de complexidade dos problemas de ambiguidade encontrados em seus conteúdos. Então, avaliou-se a qualidade dos resultados obtidos em cada um destes casos. Nesta seção, serão exibidos alguns dos documentos analisados e discutidos os respectivos resultados obtidos. Em alguns exemplos apresentados neste capítulo, são utilizados os

símbolos  e  para representar os termos que foram reconhecidos pelo *parser* como nomes de lugar e que, respectivamente, foram aceitos ou descartados após avaliação do grau de confiança. Já a utilização do símbolo ⋮ significa que, no local em que este aparece, um trecho do documento (não relevante para o exemplo descrito) foi ocultado.

A Figura 5.1 mostra um documento extraído da Web - uma reportagem sobre a festa de encerramento do carnaval da cidade do *Recife*, capital do estado de *Pernambuco*. Este exemplo retrata um caso pouco complexo, onde foi possível resolver com sucesso todos os casos de ambiguidade, sem ignorar nenhuma referência válida e descartando corretamente todas as referências não-válidas. No documento foram encontradas várias ocorrências dos termos *Recife* e *Olinda*, sendo associados corretamente aos respectivos municípios assim identificados. Foram determinantes para a aceitação destas referências: estarem grafadas com iniciais em maiúsculas; possuírem altos valores de CF_{CROSS} (entre 0,8055 e 0,9166); ter bons valores para CF_{TS} (*Recife* = 0,75 e *Olinda* = 0,204); e, em uma destas referências, há o relacionamento com um termo especial - *em Recife* (abaixo da fotografia). Além dessas, duas referências do tipo gentílico (*pernambucano* e *pernambucana*) foram devidamente associadas ao estado de *Pernambuco*; nestas, os valores CF_{CROSS} iguais a 0,5833 foram importantes na obtenção de valores de CR suficientes para serem aceitas.

Ainda no exemplo da Figura 5.1 cinco referências a três lugares não atingiram o valor mínimo de confiança e foram corretamente descartadas: *Marco*, *Valença* e *Centenário*. Existem dois municípios com nome *Centenário*; porém, o uso de letra inicial minúscula e o valor 0,00 para CF_{TS} e CF_{CROSS} fizeram com que tais referências não fossem aceitas. Com o nome *Valença*, há dois municípios e uma microrregião; apesar do valor de CF_{TS} ser razoável para estes municípios (0,546; e 0,105 para a microrregião) e de estar grafado com inicial em maiúscula (por se tratar de um sobrenome de pessoa), a ausência de referências cruzadas foi decisiva na eliminação destas referências. Por fim, a referência *Marco*, nome de um município, também foi descartada pelo baixo valor de CF_{TS} (igual a 0,068) – e pela inexistência de referências cruzadas.

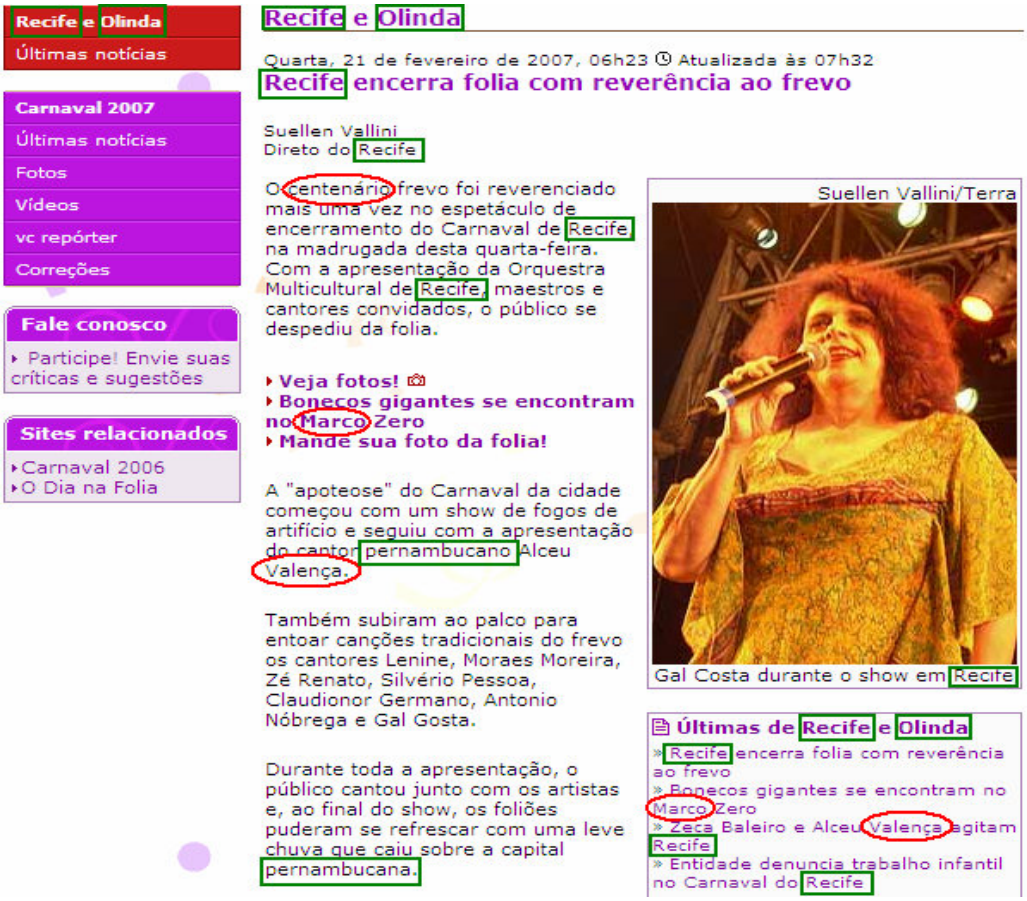


Figura 5.1 – Documento extraído da Web

A Figura 5.2 exibe outro documento capturado da Web. Este documento, obtido de um portal de turismo, apresenta algumas informações acerca de um município chamado *Telha*, localizado no estado de *Sergipe*, e reúne diversos aspectos importantes a serem observados sobre processo de detecção de referências. Um dos motivos para a escolha deste documento foi por este tratar de uma localidade com nome bastante ambíguo, cujo valor de CF_{TS} é zero. No texto, foram encontradas pelo *parser* oito referências ao município; entretanto, apenas metade destas foram aceitas após o cálculo dos valores de confiança. Dentre as referências aceitas, nota-se que todas estão associadas a algum termo especial; assim, percebe-se que o fator CF_{ST} teve papel fundamental na seleção destas referências, diante do baixo valor de CF_{TS}. Além do fator de termos especiais, foi de grande importância o valor atribuído a CF_{CROSS}, resultante da detecção das referências: *Sergipe* (estado, 3 ocorrências); *Leste*

Sergipano (mesorregião, 1 ocorrência); *Aracaju* (município, capital do estado de *Sergipe*, 1 ocorrência).

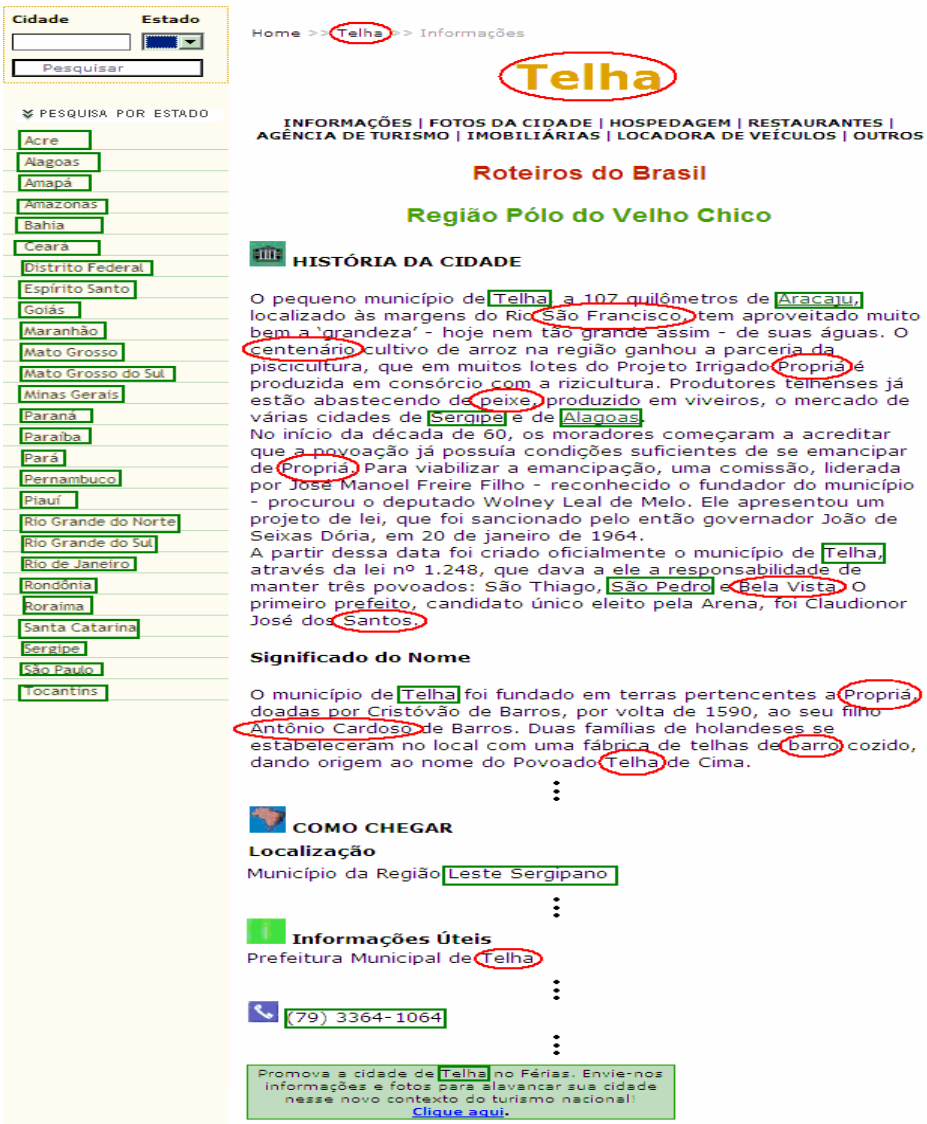


Figura 5.2 - Documento extraído da Web

Dentre as referências descartadas para o município *Telha*, as duas primeiras (posicionadas na guia de navegação da página e no título do texto) estão isoladas, ou seja, não estão inseridas em nenhum texto, fato que dificulta sua identificação, principalmente quando há um baixo valor para CF_{TS}. A terceira referência ignorada, na seção *Significado do Nome*, foi descartada corretamente, pois realmente não está relacionada ao município. Diferentemente

das referências ao município *Telha*, as referências ao estado de *Sergipe* foram corretamente detectadas, mesmo sem estarem associadas a termos especiais, visto seu alto valor de CF_{TS}. Observe que uma das referências a este estado foi originada de um número de telefone com prefixo correspondente ao estado (79), encontrado na parte inferior da página. Da mesma forma como aconteceu com algumas referências ao município *Telha*, as referências ao município *Propriá* foram ignoradas, por possuir baixo valor de CF_{TS} e por não estar associada a termos especiais. Note que o termo *Propriá* não deveria ser considerado ambíguo (deveria possuir alto valor de CF_{TS}); porém, o motor de busca acessado pelo mecanismo de captação de valores de CF_{TS} para formação dos dados de testes ignora a acentuação em seu processo de busca, fazendo com que também sejam retornados itens relacionados à palavra *Própria*, consequentemente diminuindo o valor de CF_{TS} e ocasionando ambiguidade.

Uma deficiência do mecanismo neste documento foi a detecção do município *São Pedro*, ocasionada pela ocorrência de uma referência cruzada ao estado do *Rio Grande do Norte*, bem como pelo seu valor de CF_{TS}. Outras referências foram corretamente detectadas (como a referência ao estado de *Alagoas*, encontrada na seção *História da Cidade*) e outras foram corretamente ignoradas (*São Francisco, centenário, peixe, Bela Vista, Santos, Antônio Cardoso, barro*). Um aspecto importante a ser ressaltado deste documento é a detecção de uma referência para cada estado constante no menu posicionado no lado esquerdo da página. Apesar destas referências estarem verdadeiramente relacionadas a tais estados, estas não são relevantes para o contexto da página e seria mais interessante se tivessem sido ignoradas. Páginas com esta característica podem ocasionar a construção de um contexto geográfico com excesso de localidades. Tal característica é evidenciada também em portais de notícias, de trabalhos acadêmicos, de comércio eletrônico, dentre outros. Nos portais de notícia, por exemplo, uma página possuindo conteúdo referente a uma determinada notícia dispõe de âncoras para outras páginas, contendo outras notícias (muitas vezes relacionadas ao mesmo tema, porém nem sempre ao mesmo contexto geográfico). Há ainda alguns poucos casos de sítios Web que contêm tal estrutura, porém com os links relacionados ao mesmo contexto geográfico, como acontece em alguns sítios de venda de imóveis, onde o anúncio de um imóvel possui âncoras para outros anúncios na mesma região geográfica. Outra propriedade observada em documentos HTML que ainda ocasiona falhas ao processo de detecção é o conteúdo “oculto” em alguns componentes de interface gráfica. No exemplo da Figura 5.2, há

um componente do tipo *combo-box*, posicionado no canto superior esquerdo da página, onde o usuário pode selecionar um estado para realizar uma busca dentro do portal. Apesar de não aparecer na imagem, a lista contendo tais referências está presente no código HTML e é reconhecida pelo *parser*.

Nota-se que o problema apresentado é específico de documentos HTML, não afetando outros tipo de documentos, como por exemplo, PDF e DOC. Outro ponto a se observar é que, na maioria dos casos, as referências encontradas nessas estruturas estão inseridas em um texto muito curto, ou mesmo completamente isoladas (como acontece com as referências do menu da Figura 5.2); desta forma, o seu reconhecimento é prejudicado pela ausência de alguns modificadores, como por exemplo, os de termos especiais. Assim, nem sempre acontece a detecção precipitada destas referências, uma vez que se tornam mais dependentes de um alto valor de CF_{TS}; contudo, este é o caso do documento da Figura 5.2, visto que todas as referências a estados possuem alto valor de CF_{TS}. O problema apresentado pode ser contornado com a adição do tratamento de links e âncoras nos processos de detecção e modelagem geográfica, bem como com o aprimoramento do mecanismo de detecção de referências em documentos HTML, tratando seu conteúdo de forma diferenciada de acordo com a *tag* onde está inserido.

A Figura 5.3 e a Figura 5.4 apresentam trechos de documentos HTML contendo textos fictícios, selecionados dentre os que foram criados para execução dos casos de teste. Nestes textos, utilizam-se referências a algumas localidades cujo valor de CF_{TS} é muito baixo (com valores iguais ou muito próximos a zero). Os valores exatos de CF_{TS} para cada referência utilizada nos exemplos apresentados podem ser verificados no ANEXO III. O objetivo dos testes realizados com estes documentos é avaliar o comportamento do detector de lugares diante da ocorrência de referências fortemente ambíguas, confrontando-se, no mesmo texto, ocorrências de referências válidas e inválidas para uma mesma localidade. Nestes exemplos, as referências candidatas, ou seja, as expressões que podem estar relacionadas a uma localidade geográfica, além de estarem destacadas com os símbolos  e  já conhecidos dos exemplos anteriores, podem encontrar-se numeradas (em ordem de ocorrência) para facilitar sua identificação na descrição dos resultados.

Na Figura 5.3, observam-se oito referências candidatas: *Esmeralda* (município, 2 ocorrências), *Rio Grande do Sul* (estado, 1 ocorrência), *Prata* (municípios, 3 ocorrências),

Dionísio (município, 1 ocorrência), *Uberlândia* (município e microrregião, 1 ocorrência), *MG* (sigla de estado, 1 ocorrência), *Aliança* (município, 1 ocorrência), *Ouro* (município, 2 ocorrências). Neste exemplo, apesar do contraste entre referências válidas e inválidas e da utilização de termos ambíguos, o resultado do processo de detecção foi bastante satisfatório. A referência *Esmeralda*[1] foi associada de forma correta ao município correspondente; foi importante para sua aceitação a utilização de letra inicial maiúscula, a presença de uma referência cruzada (ao estado do Rio Grande do Sul, onde o município está localizado) e por estar precedido a um termo especial (*cidade*). Por outro lado, tais características não são encontradas na referência *esmeralda*[3], ocasionando seu descarte. Já a referência *Esmeralda*[2] foi descartada apesar de seus modificadores associados. Dentre as três referências ao termo *Prata*, apenas a referência *Prata*[2] foi aceita, sendo ignoradas corretamente as demais. De forma similar à referência *Esmeralda*[1], *Prata*[2] encontra-se grafada com inicial em maiúscula, está precedida de um termo especial (*município*) e relacionada a duas referências cruzadas (*Uberlândia*, que representa um município contido na mesma microrregião; e *MG*, que representa o estado de Minas Gerais, que contêm ambos os municípios). Observa-se que, de maneira similar ao termo *Dionísio*, apenas a inicial em maiúscula não foi suficiente para a aceitação de *Prata*[1]. Além da ambiguidade ocasionada por também ser o nome de um metal, o termo *Prata* está relacionado a outro tipo de ambiguidade: a existência de duas cidades brasileiras com este nome. Assim, com a existência de referências cruzadas associadas, foi possível relacionar à localidade adequada. O termo especial *em*, precedente às duas referências *ouro*, não foi suficiente para compor seu valor de confiança acima do limiar, visto a utilização de iniciais minúsculas e a ausência de referências cruzadas, agravados pelo baixíssimo valor de CF_{TS}. Por fim, observa-se o descarte do termo *aliança*, visto sua forte ambiguidade (nome de um objeto) e a ausência de modificadores relacionados.

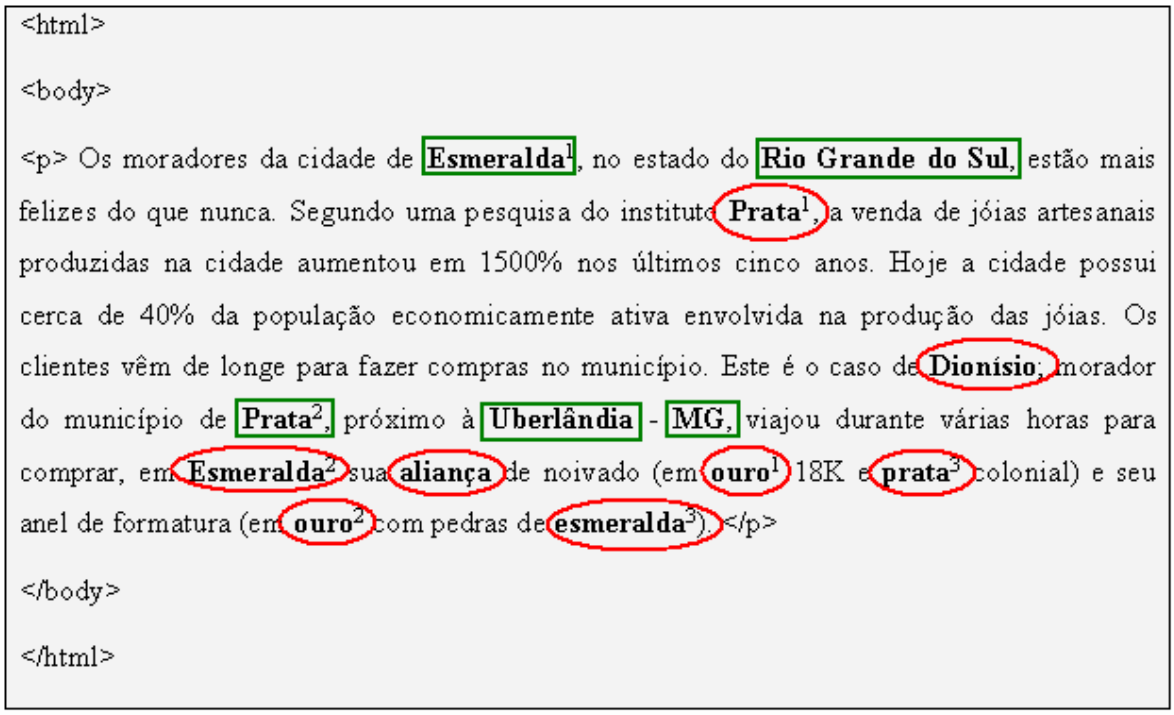


Figura 5.3 – Trecho de código HTML contendo texto fictício

Na Figura 5.4, observam-se onze referências candidatas: *Quatro Irmãos* (município, 3 ocorrências), *Noroeste Rio-Grandense* (mesorregião, 1 ocorrência), *Conquista* (município, 1 ocorrência), *Descanso* (município, 2 ocorrências), *Ouro* (município, 1 ocorrência), *Travesseiro* (município, 2 ocorrências), *Capim* (município, 3 ocorrências), *Rio Grande do Sul* (estado, 1 ocorrência), *Oeste Catarinense* (mesorregião, 1 ocorrência), *Litoral Norte* (microrregião, 1 ocorrência), *Paraíba* (estado, 1 ocorrência). Este é mais um exemplo onde o processo de detecção foi executado sem ocorrência de falhas. As referências ambíguas *Quatro Irmãos*[2], *Travesseiro*[2], *Descanso*[2] *e Capim*[3] obtiveram valores aceitáveis para seus graus de confiança, proporcionados pela grafia com letras iniciais maiúsculas, pela presença de termos especiais (respectivamente, *cidade, município, cidade e em*) e pela existência de referências cruzadas (*noroeste rio-grandense*, para *Quatro Irmãos*[2]; *Rio Grande do Sul*, para *Travesseiro*[2] e *Quatro Irmãos*[2]; *oeste catarinense*, para *Descanso*[2]; e *litoral norte* e *Paraíba*, para *Capim*[3]). *Ouro* e *Descanso*[1], apesar de grafados em maiúsculas, não obtiveram valores de confiança suficientes, pela ausência de outros modificadores. No caso deste último, note que ele também é afetado pela mesma referência cruzada de *Descanso*[2] (que foi aceita), porém em menor

intensidade devido à maior distância textual, não contribuindo o bastante para evitar seu descarte, de forma similar ao que acontece com travesseiro[1] de capim[1]. No caso destes últimos, há ainda um agravante por estarem grafados com letras iniciais minúsculas. Tem-se ainda o termo *conquista*, que não possui nenhum modificador de confiança que influencie positivamente em seu valor de confiança para compensar seu baixo valor de CF_{TS}, acarretando seu descarte. Em *Quatro irmãos*[1] e *capim*[2], o uso de iniciais minúsculas e a ausência de modificadores contribuíram para que fossem ignorados. Em *Quatro Irmãos*[3], a maior distância à referência cruzada *noroeste rio-grandense* e a ausência de termos especiais precedentes foram determinantes para seu descarte.

```
<html>

<body>

<p> Quatro irmãos[1] da cidade de Quatro Irmãos[2], no noroeste rio-grandense,
comemoram hoje mais uma grande conquista. A indústria Quatro Irmãos[3], da qual eles
compõem a sociedade, foi vencedora do prêmio Descanso[1] de Ouro, que avalia anualmente
as fábricas de produtos ligados ao sono. O principal produto da indústria, responsável por
35% de suas vendas, é o travesseiro[1] de capim[1] azul. Originário do município de
Travesseiro[2], também no estado do Rio Grande do Sul, e encontrado largamente na cidade
de Descanso[2], no oeste catarinense, este capim[2] é utilizado como matéria prima na
fabricação de diversos produtos da empresa. No próximo mês, os sócios pretendem inaugurar
uma nova fábrica em Capim[3], no litoral norte do estado da Paraíba. </p>

</body>

</html>
```

Figura 5.4 – Trecho de código HTML contendo texto fictício

Outra dificuldade observada no processo de detecção de referências foi em lidar com termos que, além de serem utilizados como nomes de lugares reconhecidos pelo protótipo, também são utilizados para identificar outras localidades fora de seu escopo. Enquadram-se nesta situação nomes de localidades não-administrativas, como por exemplo, *cachoeira* e

floresta, que também são nomes de municípios brasileiros. Além deste caso, existem ainda os nomes de lugares que são ambíguos por também serem identificadores de outras regiões administrativas, porém fora do Brasil, ou seja, fora do escopo determinado para implementação do protótipo, como por exemplo, *Colômbia, Barcelona, Coimbra, Buenos Aires*, dentre outros. Tais casos ocasionam dificuldades adicionais ao processo de detecção de referências geográficas, uma vez que suas referências podem estar associadas a modificadores de termos especiais, mesmo quando não se referem a nenhuma das localidades incluídas no escopo do protótipo, visto que, de fato, estes são lugares válidos, porém em outro contexto não conhecido pelo protótipo. Entretanto, nota-se que o problema tende a ser minimizado com a evolução do escopo do sistema, sendo este capaz de reconhecer tanto as localidades administrativas em nível global, quanto de tratar de forma diferenciada os identificadores de localidades não-administrativas.

Foram avaliados os resultados da execução do mecanismo de detecção de referências geográficas em 50 documentos previamente selecionados, com o objetivo de quantificar sua eficácia. Deste processo de avaliação, obtiveram-se os seguintes resultados aproximados:

a) 71% das referências válidas foram corretamente detectadas;
b) 29% de referências válidas foram incorretamente ignoradas;
c) 92% das referências inválidas foram corretamente ignoradas;
d) 8% das referências inválidas foram incorretamente detectadas;
e) dentre as referências que foram corretamente detectadas (item a) e que estavam relacionadas ao problema de ambiguidade onde um mesmo nome identifica mais de uma localidade, 84% foram associadas à localidade esperada;
f) dentre as referências incorretamente ignoradas (item b), 65% tiveram suas localidades detectadas por outras referências no mesmo documento;

Esses resultados podem ser descritos utilizando as métricas de revocação e precisão, de forma análoga ao método de avaliação dos resultados retornados para as consultas em motores de busca. No processo de busca, a revocação é utilizada para medir a habilidade do sistema em recuperar os documentos mais relevantes para o usuário, ou seja, mede-se o coeficiente entre a quantidade de itens relevantes que foram recuperados e total de itens relevantes existentes na base de dados. A precisão, por sua vez, mede a habilidade do sistema de manter os documentos irrelevantes fora do resultado de uma consulta, ou seja, mede-se a quantidade de

itens relevantes dentre os itens retornados para a consulta. Porém, no caso da avaliação do processo de detecção de referências geográficas, utiliza-se a revocação para mensurar a quantidade de referências válidas que foram detectadas, dentre todas as referências válidas existentes; e a precisão para mensurar a quantidade de referências válidas dentre todas as referências detectadas. A Tabela 5.1 sumariza as equações utilizadas para o cálculo destas medidas em ambos os casos.

	Busca	**Detecção de Referências Geográficas**
Revocação	$\frac{\text{Número de recuperados relevantes}}{\text{Total de relevantes possíveis}}$	$\frac{\text{Número de referências válidas detectadas}}{\text{Total de referências válidas}}$
Precisão	$\frac{\text{Número de recuperados relevantes}}{\text{Total de recuperados}}$	$\frac{\text{Número referências válidas detectadas}}{\text{Total de referências detectadas}}$

Tabela 5.1 – Revocação e Precisão na detecção e busca de georreferências

Assim, a revocação obtida para os experimentos foi de 71%, correspondente ao item a. O valor calculado para a precisão, por sua vez, foi de 54%; porém, considerando-se as referências de forma isolada, ou seja, ignorando-se o caso descrito no item f. Por outro lado, considerando-se em termos de localidade (e não apenas de referências) para fins do cálculo de precisão, o valor obtido para esta medida é de 68%. Os resultados mostram que o mecanismo avaliado já apresenta uma eficácia razoável, principalmente no que se refere à revocação. Entretanto, nota-se um baixo valor para precisão, consequente da detecção incorreta de algumas referências inválidas (item d). Apesar do quantitativo associado a este tipo de falha ser consideravelmente baixo, em torno de 8%, o impacto no cálculo da precisão é elevado, uma vez que o número de referências inválidas existentes nos textos é cerca de 7,5 vezes maior do que o número de referências válidas.

Na resolução de problemas de ambiguidade onde um mesmo nome é utilizado para identificar mais de uma localidade, os resultados foram bastante satisfatórios. Note que, apesar da quantidade de referências válidas que foram descartadas (item b), grande parte das

localidades a ela associadas foram detectadas através de outras referências no mesmo documento (item f), de modo análogo ao apresentado no exemplo da Figura 5.3, onde, apesar do descarte da referência *Esmeralda*^{2}, a localidade a esta associada não foi excluída do escopo geográfico do documento, uma vez que foi detectada através da referência *Esmeralda*^{1}. Os exemplos discutidos e os resultados apresentados demonstram também a existência de lacunas a serem contornadas e melhorias a serem adicionadas ao processo apresentado. Contudo, é de se esperar que os resultados descritos apresentem considerável evolução a partir da implementação de melhorias planejadas às técnicas desenvolvidas, bem como através da incorporação de outros métodos já propostos na literatura.

5.3. Modelagem do Escopo Geográfico

Para avaliação do processo de modelagem do escopo geográfico, assume-se que os dados provenientes do processo precedente (de detecção de referências geográficas) estão corretos. Esta metodologia tem como objetivo tornar a avaliação dos resultados do processo de modelagem independente de possíveis falhas provenientes do processo anterior, visto que há forte dependência serial entre eles, ou seja, a qualidade da saída de um processo depende da qualidade da entrada que, por sua vez, é a saída processo anterior. Nesta seção, serão apresentados os resultados obtidos para alguns dos documentos analisados, bem como descritos os resultados de experimentos realizados para alguns conjuntos de localidades constituídos especificamente para execução dos testes, ou seja, que não são provenientes da detecção de referências em um documento existente.

Para exibição de algumas geotrees geradas no processo de avaliação, é utilizada uma estrutura textual endentada, onde cada nível de endentação representa um nível de profundidade na *geotree*. As propriedades de cada nó são especificadas de acordo com os símbolos: T = Tipo; W = Peso; WB = Peso Balanceado; DR = Grau de Dispersão; e R = Relevância Geográfica. A Figura 5.5 exibe um mapa da região *Nordeste do Brasil*, destacando-se (sublinhado) as localidades citadas nos exemplos descritos nesta seção, de modo a auxiliar o entendimento destes. Através do mapa é possível, por exemplo, constatar visualmente as informações acerca da distribuição espacial destas localidades.

Figura 5.5 – Localidades citadas nos exemplos da seção 5.3

A Figura 5.6 apresenta a *geotree* produzida para o documento mostrado na Figura 5.1. A árvore é composta por dois nós do tipo *D* e um nó do tipo *H*, representando as três localidades referenciadas pelo documento – *Recife* (município), *Olinda* (município) e *Pernambuco* (estado), respectivamente. Os demais nós, do tipo *I*, representam as localidades derivadas destas primeiras. O maior valor de relevância (R=12,0), foi atribuído ao nó representando o município *Recife*, uma vez que esta foi a localidade mais referenciada no texto (doze vezes). Em segundo lugar, tem-se o nó *Pernambuco* (R=4,2), cuja localidade foi detectada duas vezes no texto. A localidade *Olinda*, por sua vez, é representada pelo nó com terceiro maior valor de relevância (R=4,0), diante das quatro referências encontradas no documento.

Note que, apesar da localidade *Pernambuco* possuir menor número de referências que a localidade *Olinda*, a primeira possui valor de relevância um pouco maior em relação à segunda, visto que a localidade *Pernambuco* também é referenciada indiretamente, herdando parte dos valores de relevância de suas localidades inferiores. O nó representando a microrregião *Recife*, derivado de dois nós do tipo *D* e com grau de dispersão DR=0,237, possui relevância maior que a mesorregião *Metropolitana de Recife*, originada de um único nó e com valor de dispersão muito baixo (DR=0,003). Todavia, esta ainda possui quase o dobro

da relevância em relação à região *Nordeste do Brasil*, que também é derivada de apenas um nó.

```
Nordeste do Brasil ( T: Indirect | W: 0.23333332 | WB: 0.23333332 | DR: 0.17898037 | R: 0.2750954 )
        Pernambuco ( T: Hybrid | W: 2.1 | WB: 2.1 | DR: 1.0 | R: 4.2 )
                Metropolitana de Recife ( T: Indirect | W: 0.5 | WB: 0.5 | DR: 0.0031579272 | R: 0.501579 )
                        Recife ( T: Indirect | W: 2.0 | WB: 2.0 | DR: 0.23685293 | R: 2.4737058 )
                                Olinda ( T: Direct | W: 4.0 | WB: 2.0 | DR: 1.0 | R: 4.0 )
                                Recife ( T: Direct | W: 12.0 | WB: 6.0 | DR: 1.0 | R: 12.0 )
```

Figura 5.6 – Geotree obtida para o documento da Figura 5.1

Para elaboração dos exemplos a seguir, modificou-se gradualmente o texto da Figura 5.1, adicionando-se referências a novas localidades ou alterando-se as localidades de referências já existentes. O objetivo desta manipulação é oferecer um modo de se comparar os escopos geográficos produzidos em situações semelhantes, através do confronto entre as geotrees geradas em cada uma destas.

A Figura 5.7 mostra a *geotree* produzida para o documento da Figura 5.1 acrescido de duas referências ao município *Aracaju*, capital do estado de *Sergipe*. De acordo com a árvore, o município de *Aracaju* é considerado menos relevante ao documento em relação aos demais municípios referenciados, uma vez que possui menor quantidade de referências no texto do que os municípios de *Olinda* e *Recife*. Este dado é verificado através dos valores de relevância atribuídos ao nó: enquanto o nó *Aracaju* possui R=1,333, os demais possuem R=2,666 e R=8,0, respectivamente.

Comparando-se as informações apresentadas na Figura 5.6 e na Figura 5.7, percebe-se que, apesar de serem iguais as quantidades de referências detectadas em ambos os documentos para os municípios *Recife* e *Olinda*, os valores de relevância atribuídos aos respectivos nós não são os mesmos nas duas geotrees. Isto acontece pela aplicação da técnica de balanceamento do peso, que faz com que estas localidades sejam mais importantes para o documento relativo à Figura 5.6, uma vez que tais municípios são os únicos referenciados pelo documento. Por outro lado, o documento relativo à Figura 5.7 possui ainda as referências ao

município *Aracaju*, o que diminui sua especificidade e consequentemente a relevância de cada um dos municípios referenciados. Note que esta diferença não acontece, por exemplo, com os nós que representam o estado de *Pernambuco*, ou seja, atribui-se igual relevância à localidade em ambos os documentos, visto que este é o único estado referenciado diretamente em ambos os documentos. Outro ponto importante a ser observado na comparação entre esses escopos geográficos é que, apesar da relevância para os nós dos municípios *Recife* e *Olinda* terem diminuído com a adição do nó do município *Aracaju*, a relevância do nó referente à região *Nordeste* aumentou, uma vez que houve crescimento do número de localidades espacialmente internas à região, bem como o aumento da dispersão geográfica destas referências.

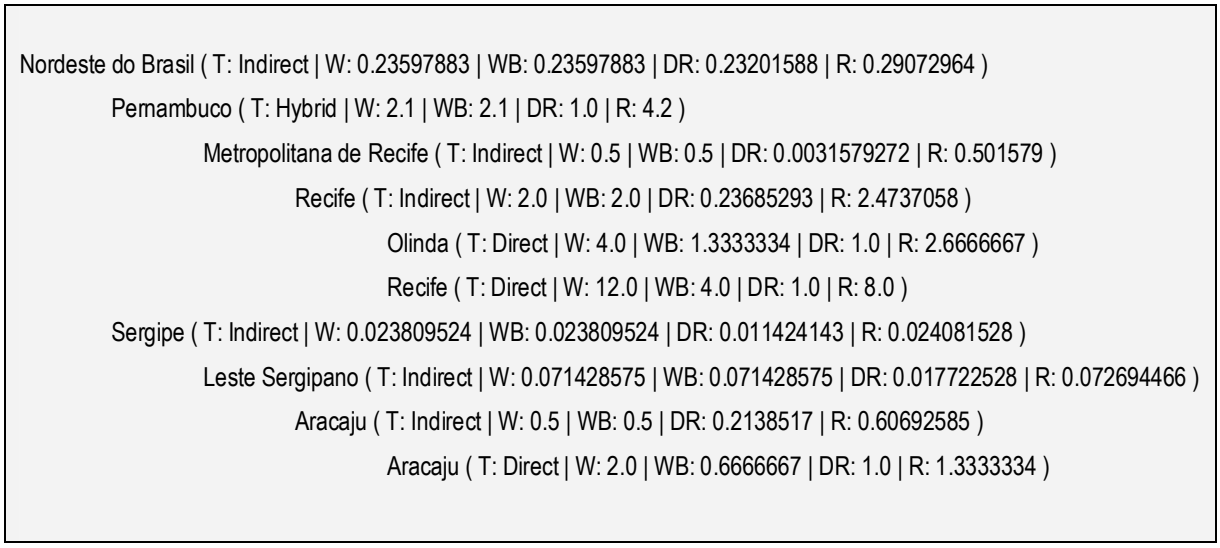


Figura 5.7 – Geotree obtida para o documento da Figura 5.1 com referências adicionais.

O documento associado à Figura 5.7 foi alterado, substituindo-se todas as referências ao município *Olinda* por referências ao município *Imperatriz*, localizado no estado do Maranhão, sendo o escopo geográfico produzido para este novo documento exibido na Figura 5.8. Nota-se que os valores dos nós *Olinda* (Figura 5.7) e Imperatriz (Figura 5.8) são iguais, uma vez que foi preservada a quantidade de referências após a substituição. De outro modo, houve uma pequena redução no valor de relevância para o nó *Pernambuco* e, mais significativamente, nos nós *Recife* (microrregião) e *Metropolitana de Recife* (mesorregião), refletindo a remoção das referências ao município *Olinda*.

Com a manutenção da quantidade de referências após as alterações do documento da Figura 5.8 em relação ao da Figura 5.7, o valor do peso (normal e balanceado) do nó *Nordeste do Brasil* permaneceu praticamente constante. Contudo, mesmo com a pouca variação no valor do peso, houve um incremento considerável no valor de relevância desse nó, ocasionado pelo aumento significativo da dispersão geográfica das localidades que compõem o documento. Através do mapa da Figura 5.5, é possível verificar de maneira visual as informações sobre a distribuição espacial destas referências. Através deste exemplo, pode-se perceber com clareza o impacto da aplicação da técnica de dispersão geográfica, fazendo com que documentos com iguais quantidades e tipos de referências possam obter valores diferentes de relevância, baseados na distribuição espacial destas localidades referenciadas.

```
Nordeste do Brasil ( T: Indirect | W: 0.23505293 | WB: 0.23505293 | DR: 0.40047023 | R: 0.32918462 )
        Pernambuco ( T: Hybrid | W: 2.075 | WB: 2.075 | DR: 1.0 | R: 4.15 )
                Metropolitana de Recife ( T: Indirect | W: 0.375 | WB: 0.375 | DR: 0.0026348266 | R: 0.37598807 )
                        Recife ( T: Indirect | W: 1.5 | WB: 1.5 | DR: 0.197619 | R: 1.7964284 )
                                Recife ( T: Direct | W: 12.0 | WB: 4.0 | DR: 1.0 | R: 8.0 )
        Sergipe ( T: Indirect | W: 0.023809524 | WB: 0.023809524 | DR: 0.011424143 | R: 0.024081528 )
                Leste Sergipano ( T: Indirect | W: 0.071428575 | WB: 0.071428575 | DR: 0.017722528 | R: 0.072694466 )
                        Aracaju ( T: Indirect | W: 0.5 | WB: 0.5 | DR: 0.2138517 | R: 0.60692585 )
                                Aracaju ( T: Direct | W: 2.0 | WB: 0.6666667 | DR: 1.0 | R: 1.3333334 )
        Maranhão ( T: Indirect | W: 0.016666668 | WB: 0.016666668 | DR: 0.0040637036 | R: 0.016734397 )
                Oeste Maranhense ( T: Indirect | W: 0.083333336 | WB: 0.083333336 | DR: 0.013446453 | R: 0.08445387 )
                        Imperatriz ( T: Indirect | W: 0.25 | WB: 0.25 | DR: 0.036036056 | R: 0.259009 )
                                Imperatriz ( T: Direct | W: 4.0 | WB: 1.3333334 | DR: 1.0 | R: 2.6666667 )
```

Figura 5.8 – Geotree obtida para o documento da Figura 5.7 alterado.

Os resultados obtidos com o mecanismo de modelagem do escopo geográfico foram bastante satisfatórios, sendo produzidas as informações esperadas nos diferentes casos avaliados. Com a geração da *geotree* de forma adequada, os documentos podem então ser indexados espacialmente, segundo as localidades representadas por seus nós, com seus

respectivos valores de relevância pré-calculados, proporcionando maior eficiência ao processo de busca destes documentos.

5.4. Execução de Buscas

Para a realização de parte dos experimentos relacionados ao processo de busca, foram indexados 66.531 documentos, a partir de URLs raízes de páginas da Web com conteúdos restritos a seis temas, citados no início deste capítulo. Uma vez formado o índice, foram executadas buscas utilizando as diversas opções de especificação do espaço geográfico, e então foram avaliados os resultados quanto à coerência aos argumentos de busca apontados, à qualidade do *ranking* de relevância e ao desempenho computacional. Para algumas buscas espaço-textuais executadas, foram elaboradas outras puramente textuais que se aproximassem semanticamente das buscas originais, com o objetivo de simular a tentativa, por parte do usuário, de executar buscas com enfoque geográfico em motores de busca textuais. Nestes casos, foram avaliados de forma comparativa os resultados de cada uma destas execuções.

Em outra série de experimentos, visando avaliar o processo de busca utilizando-se as métricas de revocação e precisão, foi constituída uma amostra mais reduzida de documentos, necessária para se obter maior controle sobre o conteúdo dos documentos analisados. Esta amostra contém 100 documentos com características geográficas diversas. Uma vez que o alvo principal desta avaliação é a relevância geográfica dos documentos recuperados, configurou-se o mecanismo de elaboração do *ranking* de relevância para compor o índice final de relevância utilizando 100% para a relevância geográfica, ou seja, desconsiderando a esfera textual. Neste sentido, as métricas foram calculadas considerando apenas a relevância geográfica. Alguns dos experimentos realizados foram selecionados para serem discutidos neste trabalho. Estes experimentos, bem como os respectivos resultados obtidos, serão descritos a seguir.

EXPERIMENTO 1:

No primeiro experimento, executa-se uma busca puramente textual, utilizando-se como argumento a expressão *"turismo ecológico"*. Com isto, obtêm-se resultados relacionados aos mais diversos contextos geográficos, uma vez que este não foi delimitado. A Figura 5.9 mostra

a tela exibida pelo sistema contendo os dez resultados mais relevantes para a referida consulta. Através do conteúdo da própria tela de resultados, sem mesmo observar a descrição do escopo geográfico dos documentos, percebe-se a associação de alguns dos itens a algumas localidades, como por exemplo, os estados de *Minas Gerais (região Sudeste); Paraná (região Sul)*; *Maranhão e Ceará (região Nordeste)*.

EXPERIMENTO 2:

Para este experimento, altera-se a consulta do experimento anterior, adicionando-se um argumento espacial, através da seleção visual no mapa interativo e da escolha do operador espacial. A saber, especificou-se como localidade a região *Norte do Brasil*, e o operador espacial escolhido foi *inside*. Ou seja, a consulta submetida tem como semântica: *recuperar as páginas cujos conteúdos estejam relacionados ao turismo ecológico e cujos escopos geográficos estejam relacionados à região norte do Brasil*. A tela com os resultados mais relevantes para esta nova consulta é exibida na Figura 5.10. Neste caso, verifica-se que dentre os dez resultados mais relevantes, todos estão fortemente associados ao contexto geográfico especificado, referenciando localidades como os estados do *Pará, Amazonas, Acre* e *Amapá*, todos integrantes da *região Norte*.

Realizar uma busca com a mesma semântica da executada no experimento descrito, em um sistema de busca puramente textual, não seria viável. Visando obter-se um resultado aproximado, seria necessário que o usuário submetesse a busca especificando como argumento textual o nome de cada uma das centenas de localidades contidas na *região norte do Brasil*. Neste caso, o tamanho do argumento de busca excederia o tamanho máximo determinado, em geral, pelos sistemas de busca. Com isto, seria necessário dividir a operação em outras menores e então unir os resultados, tornando o processo muito custoso ao usuário. Outra opção mais simples seria informar como argumento textual alguma expressão do tipo: “região norte” ou “norte do brasil”. Porém, neste caso, os resultados estariam limitados às páginas contendo exatamente tais expressões, o que ocasionaria a ocultação de muitos resultados relevantes, como por exemplo, uma página que não contenha os termos especificados, mas que contenha outros como: “Amazonas”, “Manaus”. E, por outro lado, incluiria resultados não-relevantes, como por exemplo, uma página contendo a expressão “região norte mineira” (no caso de se utilizar como argumento a expressão “região norte”).

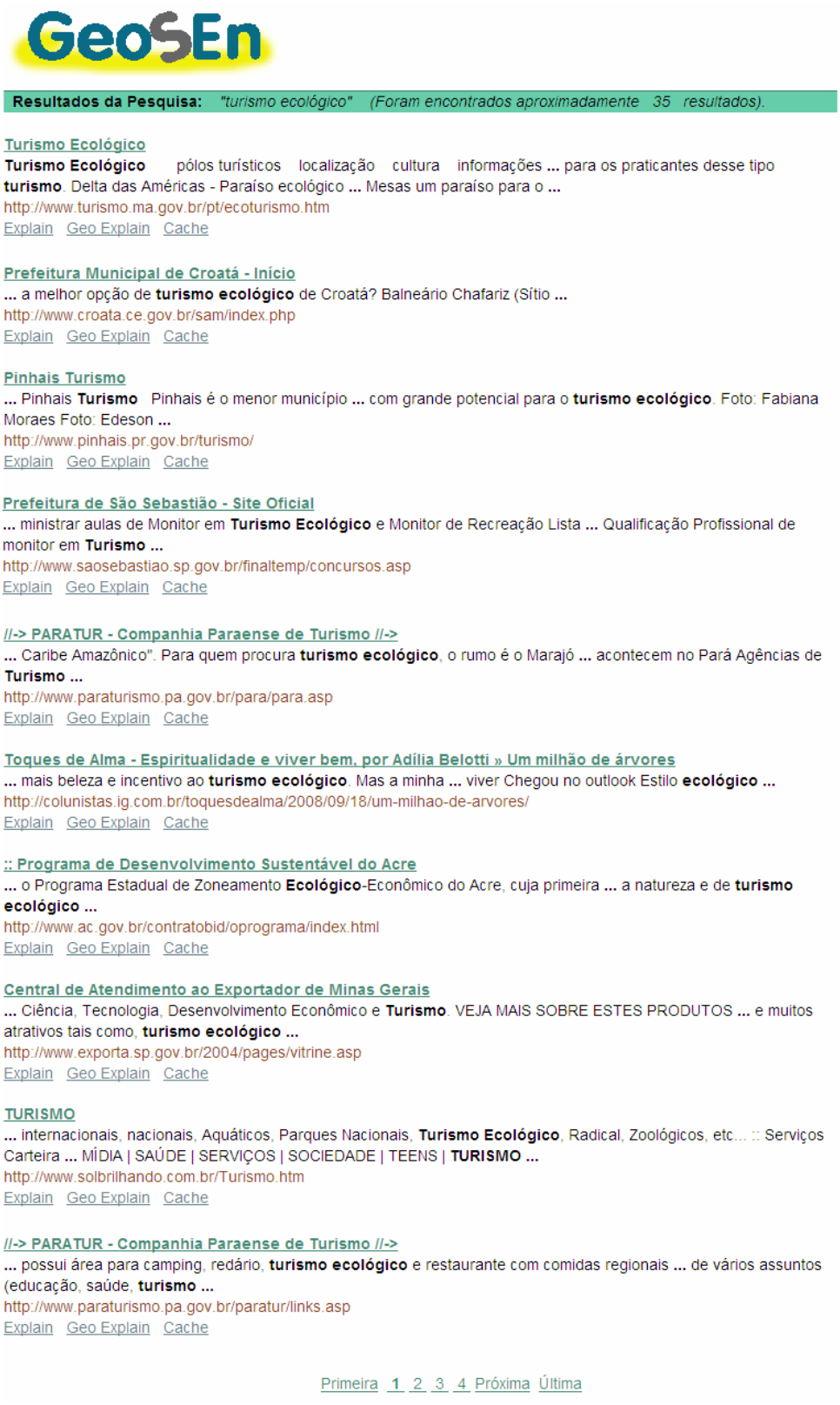
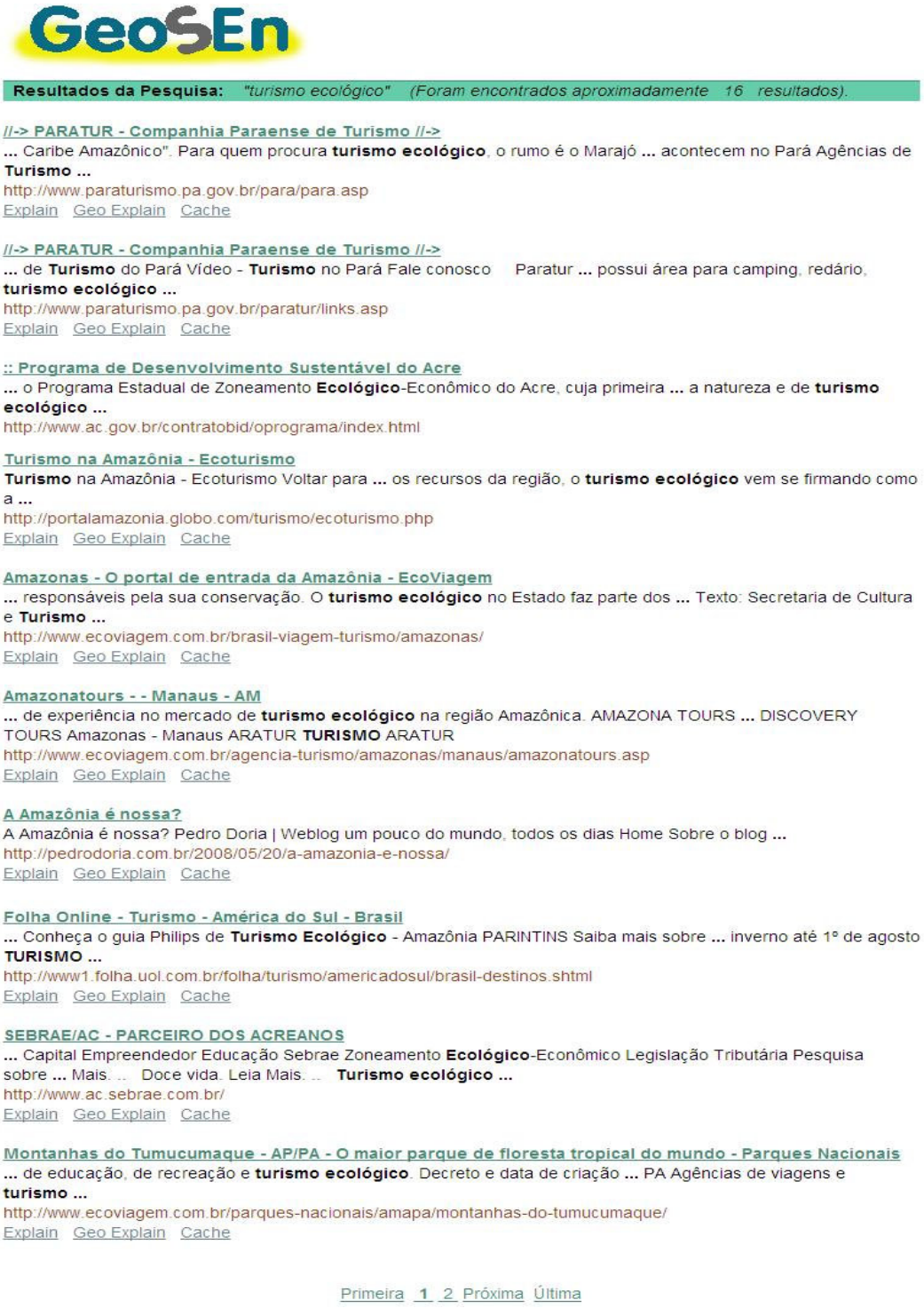


Figura 5.9 – Resultado para a busca por “turismo ecológico”


GeoSEn

Resultados da Pesquisa: *"turismo ecológico" (Foram encontrados aproximadamente 16 resultados).*

//-> PARATUR - Companhia Paraense de Turismo //->
... Caribe Amazônico". Para quem procura **turismo ecológico**, o rumo é o Marajó ... acontecem no Pará Agências de **Turismo** ...
http://www.paraturismo.pa.gov.br/para/para.asp
Explain Geo Explain Cache

//-> PARATUR - Companhia Paraense de Turismo //->
... de **Turismo** do Pará Vídeo - **Turismo** no Pará Fale conosco Paratur ... possui área para camping, redário, **turismo ecológico** ...
http://www.paraturismo.pa.gov.br/paratur/links.asp
Explain Geo Explain Cache

:: Programa de Desenvolvimento Sustentável do Acre
... o Programa Estadual de Zoneamento **Ecológico**-Econômico do Acre, cuja primeira ... a natureza e de **turismo ecológico** ...
http://www.ac.gov.br/contratobid/oprograma/index.html

Turismo na Amazônia - Ecoturismo
Turismo na Amazônia - Ecoturismo Voltar para ... os recursos da região, o **turismo ecológico** vem se firmando como a ...
http://portalamazonia.globo.com/turismo/ecoturismo.php
Explain Geo Explain Cache

Amazonas - O portal de entrada da Amazônia - EcoViagem
... responsáveis pela sua conservação. O **turismo ecológico** no Estado faz parte dos ... Texto: Secretaria de Cultura e **Turismo** ...
http://www.ecoviagem.com.br/brasil-viagem-turismo/amazonas/
Explain Geo Explain Cache

Amazonatours - - Manaus - AM
... de experiência no mercado de **turismo ecológico** na região Amazônica. AMAZONA TOURS ... DISCOVERY TOURS Amazonas - Manaus ARATUR **TURISMO** ARATUR
http://www.ecoviagem.com.br/agencia-turismo/amazonas/manaus/amazonatours.asp
Explain Geo Explain Cache

A Amazônia é nossa?
A Amazônia é nossa? Pedro Doria | Weblog um pouco do mundo, todos os dias Home Sobre o blog ...
http://pedrodoria.com.br/2008/05/20/a-amazonia-e-nossa/
Explain Geo Explain Cache

Folha Online - Turismo - América do Sul - Brasil
... Conheça o guia Philips de **Turismo Ecológico** - Amazônia PARINTINS Saiba mais sobre ... inverno até 1º de agosto **TURISMO** ...
http://www1.folha.uol.com.br/folha/turismo/americadosul/brasil-destinos.shtml
Explain Geo Explain Cache

SEBRAE/AC - PARCEIRO DOS ACREANOS
... Capital Empreendedor Educação Sebrae Zoneamento **Ecológico**-Econômico Legislação Tributária Pesquisa sobre ... Mais. .. Doce vida. Leia Mais. .. **Turismo ecológico** ...
http://www.ac.sebrae.com.br/
Explain Geo Explain Cache

Montanhas do Tumucumaque - AP/PA - O maior parque de floresta tropical do mundo - Parques Nacionais
... de educação, de recreação e **turismo ecológico**. Decreto e data de criação ... PA Agências de viagens e **turismo** ...
http://www.ecoviagem.com.br/parques-nacionais/amapa/montanhas-do-tumucumaque/
Explain Geo Explain Cache

Primeira 1 2 Próxima Última

Figura 5.10 - Resultado para a busca por "turismo ecológico" na região Norte

EXPERIMENTO 3:

No terceiro experimento, avalia-se a qualidade dos resultados para uma busca utilizando o operador espacial de distância. Neste, a consulta submetida tem como significado: *recuperar as páginas cujos conteúdos estejam relacionados a concursos e cujos escopos geográficos estejam relacionados a alguma localidade com distância menor ou igual a 300Km da cidade de Campina Grande*. Para isto, especifica-se como argumento textual o termo *concurso*, seleciona-se o operador espacial *distance*, e utiliza-se o valor 300Km como parâmetro do operador. No domínio espacial, seleciona-se o município de *Campina Grande*.

Os resultados mais relevantes para busca executada no terceiro experimento estão exibidos na Figura 5.11. Dentre estes, tem-se, por exemplo:

- os 1°, 2°, 3°, 5°, 8° e 9° resultados, contendo referência à cidade de *Recife-Pe*, localizado a menos de 200 Km de *Campina Grande*;
- os 4° e 5° resultados, relacionados à cidade de *João Pessoa*, capital do estado da Paraíba, onde também está localizado o município de *Campina Grande* (e distante menos de 130 Km deste);
- o 6° resultado, contendo referências às cidades de *Araruna*, *Cacimba de Dentro* e *Campo de Santana*, todas localizadas no *agreste paraibano* - mesma mesorregião onde está inserido o município de *Campina Grande* - e todos localizados a menos de 150 Km do município;
- o 7° resultado, referenciando o próprio município de *Campina Grande* e outros também do estado da *Paraíba*, localizados a menos de 200 Km do município especificado, a saber, *João Pessoa*, *Cabedelo*, e *Patos*;
- o 10° resultado, relacionado à cidade de *Olinda*, cuja distância à *Campina Grande* também é inferior a 200 Km.

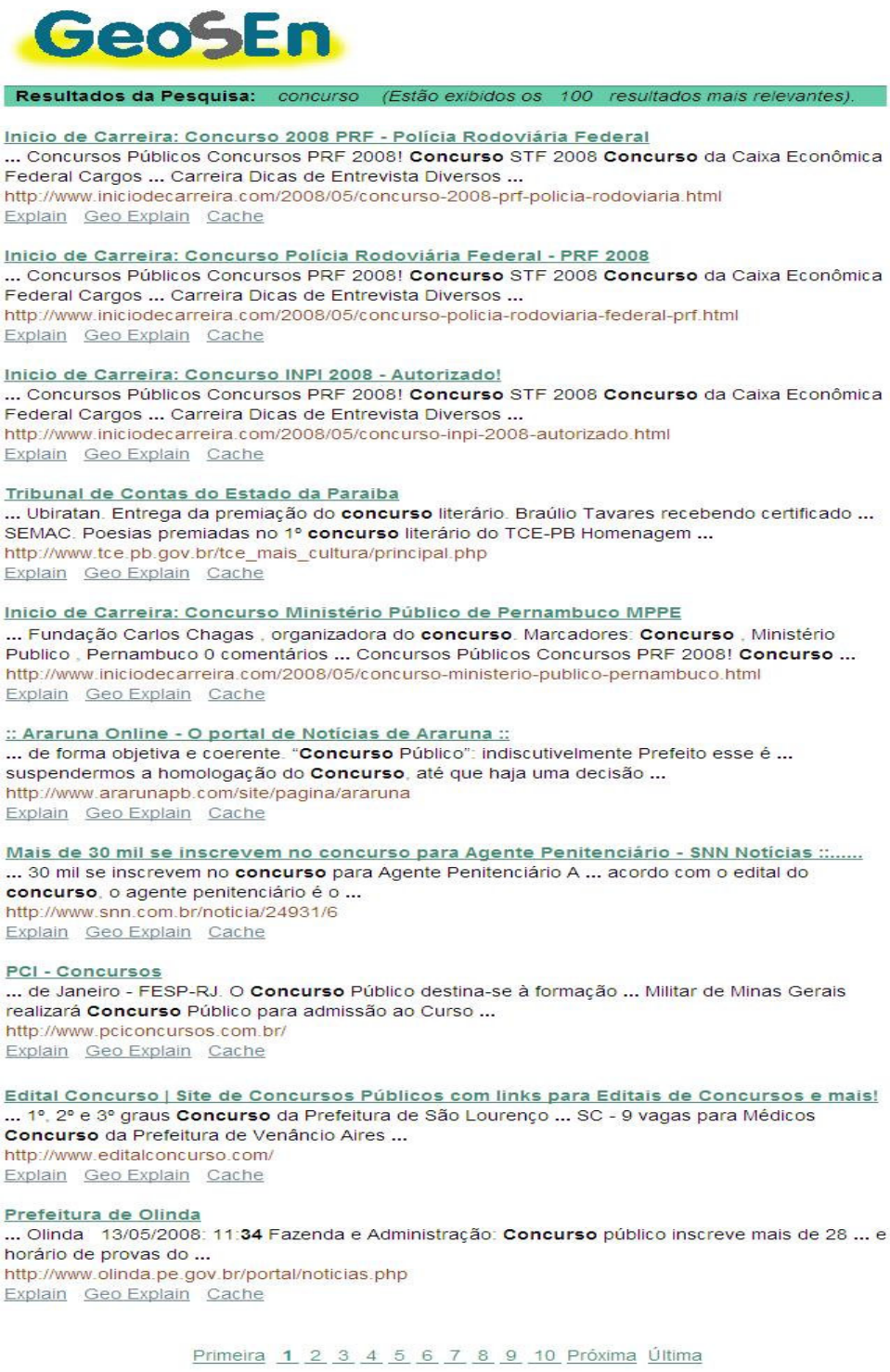


Figura 5.11 - Resultado para a busca por concurso distante até 300 Km da cidade de Campina Grande

EXPERIMENTO 4:

Para este, a consulta do experimento anterior é um pouco modificada, alterando-se o operador de *distance* para *not + distance*, mudando a semântica da busca para: *recuperar as páginas cujos conteúdos estejam relacionados a concursos e cujos escopos geográficos não estejam relacionados a alguma localidade com distância menor ou igual a 300Km da cidade de Campina Grande*. Os resultados mais relevantes obtidos para essa nova consulta são apresentados na Figura 5.12. Como esperado, nenhum resultado está relacionado a alguma localidade num raio de 300 Km de Campina Grande. Dentre as localidades relacionadas aos resultados apresentados, citam-se como exemplo, os estados de *São Paulo*, *Goiás* e *Minas Gerais*.

Realizar as pesquisas dos últimos dois experimentos, utilizando-se um motor de busca textual, torna-se ainda mais complicado do que os casos anteriores, pois, mesmo para submeter a pesquisa para todas as localidades possíveis, seria necessário que o usuário conhecesse todas as localidades existentes naquele raio de distância. No último experimento apresentado, foi avaliado o operador de negação. Com este, é possível especificar as localidades que se deseje excluir dos resultados da busca. Este operador é muito útil em alguns tipos de pesquisa, onde os resultados retornados estão fortemente relacionados a um conjunto de localidades. Por exemplo, ao se efetuar uma busca textual na base de dados indexada para estes experimentos, utilizando-se como argumento o termo *carnaval*, os resultados mais relevantes estão, em sua maioria, relacionados ao estado da Bahia e/ou à sua capital, a cidade de Salvador, pelo fato desta festa ser muito importante para a cultura desta região, e por esta promover fortemente o turismo no período de realização desta festa. Porém, pode ser de interesse do usuário pesquisar sobre o carnaval ignorando a sua relação com a *Bahia*, ou seja, excluindo dos resultados as páginas cujos escopos referenciem o estado. Para isto, basta que se selecione o estado como argumento espacial e *not + inside* como argumento espacial.

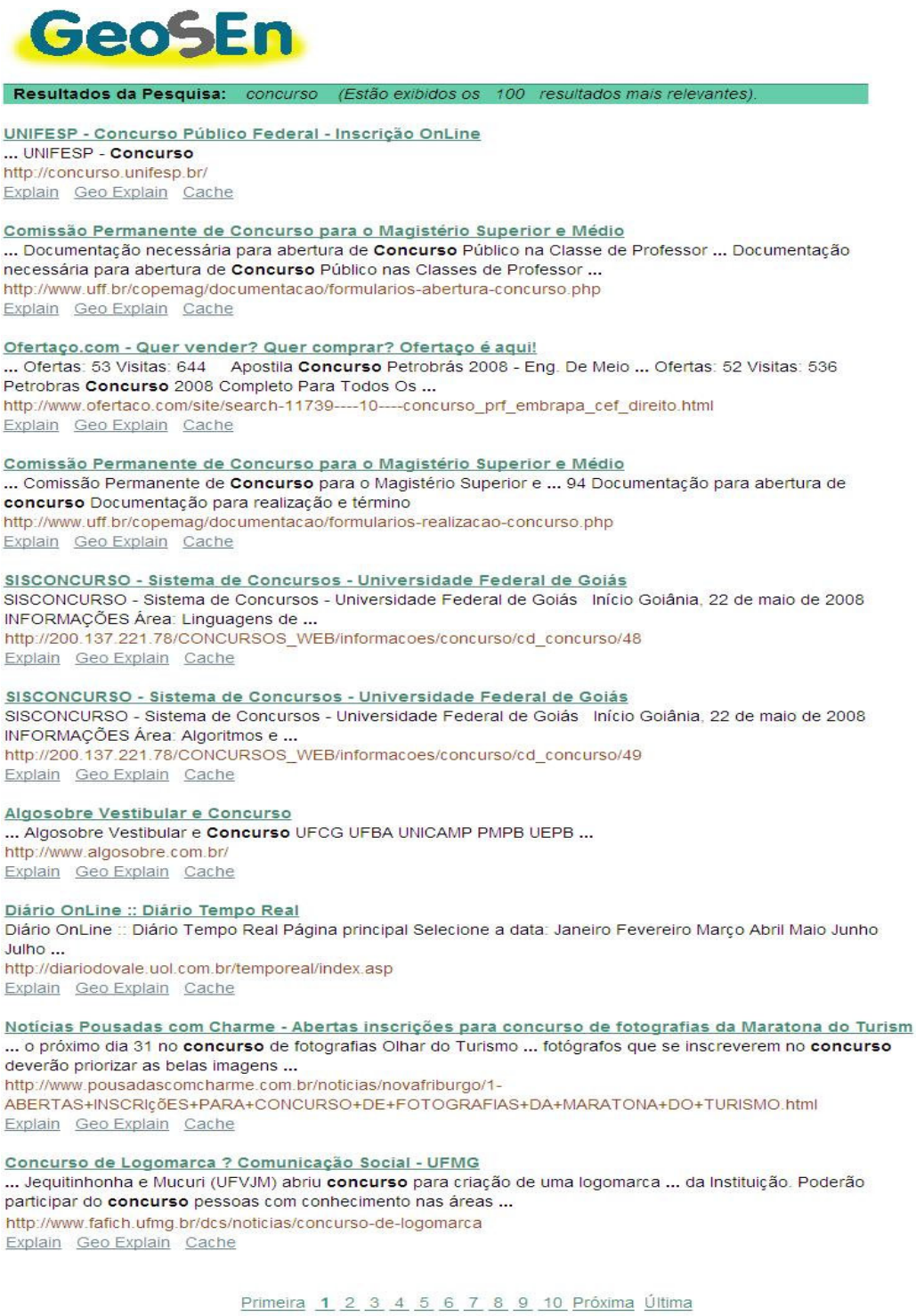


Figura 5.12 - Resultado para a busca por concurso distante pelo menos 300 Km da cidade de Campina Grande

EXPERIMENTO 5:

No quinto e último experimento desta série, avalia-se o processo de busca especificando-se visualmente o argumento geográfico através da ferramenta de seleção retangular. Objetivo desta consulta é *retornar as páginas cujos conteúdos estejam relacionados com obras de prefeituras e cujos escopos geográficos estejam relacionados com a região delimitada no mapa*. O argumento textual utilizado para esta consulta foi *obras prefeitura* e o operador espacial escolhido foi *inside*. A Figura 5.13 exibe a região retangular especificada. Esta região intersecta parte dos estados do *Rio Grande do Norte* e da *Paraíba*, incluindo as capitais *Natal* e *João Pessoa*.

A Figura 5.14 mostra os dez primeiros resultados obtidos para a busca do quinto experimento, que se mostram bastante relevantes no que se diz respeito ao contexto geográfico. Em meio aos resultados retornados, encontram-se páginas relacionadas a alguns municípios localizados na região especificada, como por exemplo, Natal-Rn (1°), João Pessoa-Pb (2°, 3°, 9° e 10°), Lucena-Pb (2° e 10°), Cabedelo-Pb (2°, 4° e 9°), Goianinha-Rn (7°) Caaporã (10°), dentre outros. Neste experimento, observa-se a ocorrência de um resultado não-relevante (6°), alí presente por supostamente referenciar a cidade de Natal-Rn. No entanto, tal referência foi associada ao escopo da página incorretamente, pois o termo *Natal* existente em seu texto foi empregado com outro sentido. Ou seja, no caso apresentado, o mecanismo de busca funcionou de maneira adequada, porém transmitindo falhas provenientes de um processo anterior.

Figura 5.13 – Seleção retangular para a busca por *obras prefeitura*

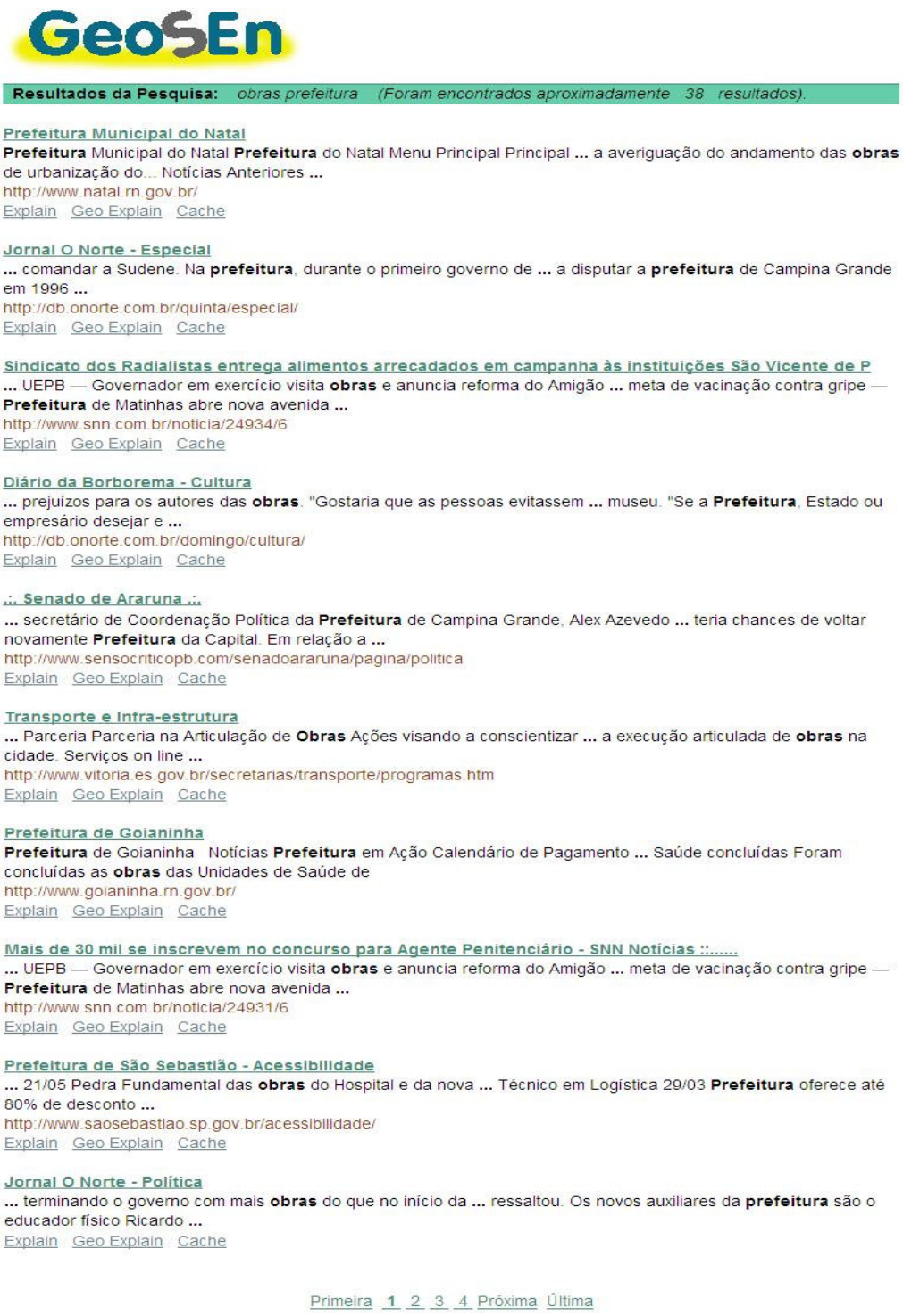


Figura 5.14 - Resultado para a busca por *obras prefeitura* na região delimitada por seleção retangular

O mecanismo de busca do protótipo foi avaliado também através das métricas revocação e precisão, já apresentadas anteriormente. Em uma nova série de experimentos, executou-se um conjunto de 50 buscas à base completa (contendo os 66.531 documentos), a exemplo dos experimentos mostrados anteriormente nesta seção, variando-se tanto os argumentos textuais quanto os espaciais. Para cada busca executada, avaliaram-se manualmente as informações geográficas de cada item do resultado (utilizando-se a função *geo explain*, disponível na tela de resultados), com o objetivo de mensurar a sua coerência com os argumentos de busca especificados.

O procedimento descrito é capaz de avaliar o processo de busca de forma isolada, ou seja, considerando que os dados gerados para os escopos geográficos estão sempre corretos. Com estes experimentos, não foram detectadas falhas no processo de recuperação dos documentos, ou seja, ao se conferir o escopo geográfico de cada item do resultado, observa-se que nenhum resultado da busca é irrelevante em relação ao argumento espacial informado, e também que nenhum item relevante a tal argumento foi desprezado pela consulta.

Nota-se que os referidos valores de revocação e precisão não refletem a qualidade do processo de recuperação de informação quando considerado de forma ampla, ou seja, desde o *parsing* até a execução da busca. Com isto, realizou-se ainda outra série de experimentos, onde se assumiu que os dados exibidos no *geo explain* podem estar incorretos, ou seja, foram considerados os erros provenientes dos processos anteriores. Nestes experimentos, não foram observados apenas os dados do *geo explain*, mas também o conteúdo de cada documento retornado. Com este procedimento, foi possível estimar o valor de precisão, não sendo viável, porém, estimar o valor de revocação, visto que para este seria necessário conhecer o conteúdo de todos os documentos indexados. Os valores mínimo, médio e máximo obtidos para precisão foram, respectivamente, 66%, 83% e 94%.

Em uma última série de experimentos, utilizou-se uma base de documentos com quantidade reduzida de itens, visando se obter um maior controle sobre o conteúdo dos mesmos, de forma a estimar também o valor de revocação, além do valor de precisão. Esta base é constituída por cem documentos, extraídos da Folha On-Line. O conteúdo de cada página indexada é uma matéria jornalística diferente, contendo uma ou mais ocorrências para a expressão “meio ambiente”. Deste modo, fixou-se tal expressão como argumento textual de busca para os experimentos realizados, variando-se apenas o argumento espacial, com o

objetivo de dar maior ênfase à perspectiva espacial do processo de busca. Cada documento foi avaliado também de forma manual, no que se refere ao seu contexto geográfico, de modo a se mensurar a qualidade das informações de escopo geográfico geradas automaticamente.

O ANEXO IV mostra as localidades referenciadas pelos documentos e a quantidade de páginas onde uma determinada localidade é referenciada. As localidades mais referenciadas nos documentos foram São Paulo (estado), São Paulo (cidade), Porto Alegre, Mato Grosso, Rondônia, Pará, Minas Gerais, Embu, Rio de Janeiro (estado), Rio de Janeiro (cidade), Paraná, Amazonas, Maranhão, Mato Grosso do Sul e Tocantins, respectivamente, com 96, 94, 56, 36, 34, 32, 24, 12, 20, 20, 16, 14, 14, 12 e 10 ocorrências. Nota-se que as três localidades mais freq uentes possuem quantidades de ocorrência bastante superiores às demais.

No caso do estado de São Paulo, por exemplo, foram encontradas referências em todos os documentos analisados. Uma vez que todos os documentos foram extraídos da Folha On-Line, do mesmo produtor do Jornal *Folha de S. Paulo*, verificou-se, em todas as páginas acessadas, ocorrências do termo *São Paulo*, porém, muitas vezes relacionado ao nome do Jornal, e não propriamente ao estado. Durante o processo de avaliação manual, constatou-se que apenas treze destes documentos possuem conteúdo relacionado ao estado de São Paulo. Já no caso da cidade de Porto Alegre, por exemplo, todas as referências foram encontradas em um mesmo texto publicitário que constava nestas páginas no momento da execução do processo de *crawling*. As referências à cidade de Embu, por sua vez, foram encontradas na seção de comentários dos leitores, no final de cada matéria.

Nestes últimos experimentos, utilizou-se um critério mais subjetivo na análise da relevância geográfica dos documentos, baseado em investigações manuais. Por exemplo, um documento contendo uma reportagem sobre o desmatamento no estado do Mato Grosso, não foi considerado relevante para o estado de São Paulo, por conter ocorrências a este pelo motivo exposto anteriormente. Realizou-se um conjunto de 50 buscas, variando-se apenas o argumento espacial, procurando-se distribuir as localidades informadas de modo a contemplar todo o território nacional. Os resultados obtidos para estes experimentos estão exibidos na Tabela 5.2.

	Revocação	**Precisão**
Mínimo	65%	62%
Médio	79%	81%
Máximo	88%	91%

Tabela 5.2 – Resultados para a terceira série de experimentos do processo de busca

5.5. Escalabilidade

O Nutch é um sistema de busca escalável, capaz de indexar da ordem de milhões de páginas por dia por CPU, e de executar de forma distribuída, utilizando grades computacionais. No intuito de explorar essas características do Nutch, os plugins do GeoSEn mantêm a mesma estrutura de indexação textual na manipulação de suas informações espaciais, sendo necessário uma quantidade pequena de operações espaciais em tempo de execução de buscas, conforme discutido na seção 4.6. A quantidade de itens armazenados nos novos campos criados no índice são proporcionais à quantidade de localidades existentes no escopo geográfico do documento. Mesmo em casos bastante raros, onde a quantidade de itens contidos no escopo pode ser relativamente grande (i.e., mais de 50 elementos), isso não se mostra um empecilho a escalabilidade do sistema, uma vez que esse índice suporta campos muito maiores, como por exemplo, o campo contendo os termos que compõem o texto do documento.

No que se refere ao processo de coleta das páginas (*crawling*), o GeoSEN adiciona uma quantidade considerável de novas funções ao mecanismo do Nutch. Diante disto, tais funcionalidades foram implementadas de forma a diminuir o *overhead* na análise dos documentos, como por exemplo, utilizando estruturas de indexação em memória para eliminação de acesso à banco. Com isto, obteve-se um tempo de processamento de 3 a 5 vezes maior que o tempo original (sem o processamento dessas novas funcionalidades), o que não compromete a escalabilidade do sistema como um todo. Para o cálculo do ranking de relevância da forma como foi proposta, o tempo de resposta foi aceitável diante do tempo total esperado para um sistema de busca Web. Entretanto, é possível que em testes com coleções maiores de documentos indexados apontem necessidades de melhorias neste ponto,

principalmente pela necessidade de normalização dos valores de relevância geográficos imposta pelo modelo proposto.

5.6. Conclusão

Neste capítulo foram descritos os métodos adotados e os resultados obtidos nos experimentos realizados para validação do protótipo desenvolvido. Através dos experimentos apresentados, é possível constatar que a maioria das falhas observadas são decorrentes de lacunas no processo de detecção de referências geográficas, sendo refletidas nos processos subsequentes. Deste modo, identifica-se este módulo como o principal alvo de melhorias a serem desenvolvidas em trabalhos futuros.

Os processos de modelagem do escopo geográfico e de busca apresentaram um ótimo desempenho quando avaliados de maneira isolada. Quando o processo de busca é analisado de forma ampla, ou seja, observando-se a qualidade dos resultados obtidos para a execução das buscas considerando as falhas decorrentes das etapas anteriores, obteve-se um valor médio de revocação e precisão de aproximadamente 79% e 81%, respectivamente. Estima-se que, com a introdução de melhorias planejadas ao processo de detecção, estes valores médios possam ultrapassar os 90%.

Capítulo 6
Conclusão

Neste trabalho foram expostos os principais fundamentos da recuperação de informação e dos motores de busca para Web, descrevendo-se os aspectos arquiteturais destes sistemas e os modelos clássicos de RI que são utilizados como base para os modelos adotados nos principais sistemas de busca modernos. Descreveu-se ainda o estado da arte em recuperação de informação geográfica, onde foram apresentados as principais contribuições e os respectivos projetos de pesquisa relacionados ao tema.

Foram descritos os métodos e técnicas elaboradas para os processos de reconhecimento de referências geográficas em documentos da Web, de modelagem do escopo geográfico destes documentos, e de execução de busca considerando os aspectos textuais e espaciais. Apresentou-se o protótipo desenvolvido para validação dos métodos propostos. Do protótipo, foram discutidos os aspectos arquiteturais, os principais algoritmos implementados, e os experimentos realizados para avaliação funcional.

6.1. Contribuições

Foi proposto um modelo de implementação baseado na extensão de um sistema de *crawling* e busca para a Web, com código fonte aberto e arquitetura orientada a plugins. No projeto elaborado, coletam-se dados de fontes externas e mantém-se uma base de dados geográficos utilizando um SGBD de distribuição livre e com suporte espacial.

Sugeriu-se um método de detecção de referências geográficas em documentos Web, baseado em um conjunto de heurísticas. No método proposto, foram introduzidos os conceitos

de *fator de confiança* e de *modificadores de confiança*, que visam mensurar a probabilidade de uma referência geográfica detectada ser uma referência válida e de estar associada à localidade correta, em processos de eliminação de ambiguidades.

Foi apresentado um mecanismo inédito de atribuição de escopo geográfico às páginas da Web, capaz de realizar suas análises e calcular valores de relevância com base na dispersão geográfica das localidades referenciadas nos documentos, e em estatísticas extraídas da hierarquia de localidades explorada. Neste mecanismo, utiliza-se uma técnica intitulada de *expansão do georreferenciamento*, onde novas localidades são adicionadas ao escopo do documento a partir das localidades referenciadas diretamente.

Descreveram-se os mecanismos de indexação e busca desenvolvidos. No mecanismo de indexação, os documentos são indexados segundo as localidades pertencentes ao seu escopo geográfico (tanto as detectadas diretamente quanto as derivadas através do método de expansão), com valores de relevância associados. Com a adição de novas localidades referenciadas indiretamente e com o cálculo dos valores de relevância em tempo de indexação, obtém-se uma redução considerável no custo computacional em tempo de realização de busca. No mecanismo de busca, utiliza-se uma interface multi-modo, onde é possível especificar textualmente ou visualmente a região explorada no processamento da busca.

6.2. Trabalhos Futuros

O GeoSEn já possui diversas funções agregadas. Entretanto, por se tratar de um protótipo, este ainda tem várias limitações funcionais, das quais destacam-se:

- localidades reconhecidas restritas às regiões administrativas brasileiras, contendo a hierarquia município → microrregião → mesorregião → estado → região;
- quantidade reduzida de documentos indexados (da ordem de dezenas de milhares de documentos);
- processo de *crawling* limitado a URLs do domínio ".br" e documentos em formato HTML;
- processo de *parsing* restrito a conteúdos em português;

São diversas as possibilidades de continuidade deste trabalho. Projetos de extensão visam a eliminação das restrições supracitadas, o aprimoramento das técnicas e métodos desenvolvidos e a introdução de novas funcionalidades. A seguir, estão relacionadas as principais contribuições visualizadas, onde algumas delas estão em fase de desenvolvimento.

Internacionalização. Algumas limitações apresentadas serão eliminadas com a extensão da ferramenta para manipulação de localizações em nível global e com sua internacionalização, ou seja, com a adição da capacidade de manipular documentos escritos em múltiplos idiomas.

Adição de heurísticas. No que se diz respeito ao processo de detecção de referências geográficas, o refinamento dos algoritmos utilizados podem ser acompanhados da incorporação de outras heurísticas já apresentadas na literatura, como por exemplo, as descritas por Silva et al [38]. Dentre os pontos passíveis de extração de tais referências, a exploração do *whois* ainda não foi implementada, por não ter sido encontrado até o presente momento nenhum *Web Service* que fornecesse livre acesso aos dados dos registros brasileiros. Entretanto, como existem serviços disponíveis para domínios internacionais, planeja-se a implementação deste módulo durante o processo de internacionalização.

Heurísticas que visam o aprimoramento do processo de *parsing* de documentos HTML, tratando seu conteúdo de forma diferenciada de acordo com a *tag* onde está inserido, mostra-se importante tanto para o aperfeiçoamento do mecanismo de detecção de referências quanto para o de modelagem do escopo geográfico. Por exemplo, é possível atribuir diferentes valores de importância a uma referência geográfica de acordo com a formatação com a qual esta ocorre no texto. Ainda, com o tratamento diferenciado do conteúdo das âncoras de um documento, torna-se possível contornar determinadas falhas relacionadas ao processo de detecção de georreferências, como por exemplo, as descritas na seção 5.2.

O detector de termos especiais (vide seção 4.2.2) está sendo adaptado para ser capaz de reconhecer expressões (e.g., "zona norte de", "região metropolitana de"), visto que atualmente este reconhece apenas termos isolados; e ainda a utilização de termos especiais de forma combinada, especificando regras para que a confiança possa ser modificada com a ocorrência destas. Exemplo de uma regra:

- Termo especial $ST_A \rightarrow CF_A = x$
- Termo especial $ST_B \rightarrow CF_B = y$
- Termo especial ST_A *seguido* de $ST_B \rightarrow CF_{AB} = z$

Exploração da estrutura da web. Outro aspecto importante a ser investido é a exploração da estrutura de ligação da Web, a exemplo do apresentado por Martins et al [39]. Tal característica é bastante útil no refinamento do escopo atribuído às páginas e também na identificação dos locais de interesse por uma determinada página.

Novos operadores espaciais. Tratando-se do processo de execução de buscas, pretende-se disponibilizar uma quantidade maior de operações espaciais (atualmente estão disponíveis as operações de continência e distância), bem como a possibilitar a construção de expressões lógicas na especificação das localidades.

Balanceamento dinâmico. Adicionalmente, se mostra interessante a inclusão de recursos como o apresentado por Markowetz [51], que permite ao usuário, em tempo de execução, fazer o balanceamento entre as perspectivas textual e espacial na elaboração do *ranking* de relevância. Atualmente este balanceamento é configurado pelo administrador do sistema.

Manipulação de imagens. Está em fase de desenvolvimento no GeoSEn um processo de *crawling*, indexação e busca de imagens, similar ao existente para os documentos da Web. Tais imagens são extraídas dos documentos analisados em formato HTML. Cada imagem coletada passa por um processo de *parsing*, com o objetivo de indexá-las textualmente e espacialmente, através da análise de alguns atributos, como seu nome e os metadados contidos no cabeçalho de alguns formatos de arquivo, como o JPEG.

As imagens coletadas são indexadas segundo suas palavras-chave (extraídas dos atributos mencionados) e escopo geográfico. Para detecção do escopo geográfico, os elementos mais importantes são seu nome e, quando disponível, a informação de posicionamento geográfico contida nos metadados de alguns arquivos. Tais informações são pouco freq uentes nas fotos atualmente obtidas na Web, todavia, existe uma tendência de aumento de sua disponibilidade, devido ao crescimento da quantidade de câmeras fotográficas com GPS embutido, principalmente daquelas acopladas a aparelhos de telefonia móvel e computadores portáteis.

Além das informações obtidas diretamente de atributos das imagens, o escopo geográfico destas é derivado também do escopo geográfico das páginas que as contêm, podendo estas herdá-los de forma parcial ou total. A herança parcial do escopo se faz necessária, por existirem páginas da Web que contêm várias imagens e que também estão associadas a um escopo geográfico múltiplo, ocasionando ambiguidade no delineamento do escopo de cada uma destas imagens. Por exemplo, considere uma página contendo imagens de três cidades -A, B e C - e tendo seu escopo associado às mesmas três cidades. Deste modo, se cada uma destas imagens herdasse o escopo total da página (i.e., escopo de A = escopo de B = escopo de C = A, B, C), resultaria em uma modelagem inconsistente, visto que o correto deve ser associar cada uma delas ao seu respectivo escopo (escopo de A = A, escopo de B = B e escopo de C = C). Nestes casos, utiliza-se uma técnica de associação por proximidade, onde cada imagem herda o escopo geográfico relacionado às referências detectadas mais próximas de si. Por fim, tem-se que os processos de expansão do georreferenciamento e de cálculo da dispersão geográfica, também são aplicados à indexação das imagens, da mesma forma como são executados para os documentos.

Utilização do mapa para saída de informações. Através do mapa interativo implementado do GeoSEn é possível especificar os parâmetros espaciais da busca, ou seja, este é utilizado para a entrada de dados no sistema. No entanto, a utilização deste mapa para saída de informações também se mostra bastante útil, sendo tal implementação planejada para as atividades futuras de desenvolvimento. Nesta abordagem é possível, por exemplo, executar uma busca apenas com parâmetros textuais e visualizar no mapa o escopo geográfico de cada item do resultado, individualmente ou de forma coletiva. Aliado à capacidade de identificar os locais de interesse de uma página, este recurso pode se mostra interessante, por exemplo, para realização de pesquisas de mercado. Para isto, o usuário pode informar a página de uma empresa e visualizar no mapa os locais de interesse.

Referências Bibliográficas

[1] C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. J. van Kreveld, R. Weibel. *Spatial information retrieval and geographical ontologies an overview of the SPIRIT project*. SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 387-388. 2002.

[2] L. Gravano. GeoSearch: *A Geographically-Aware Search Engine*, 2003. www.cs.columbia.edu/~gravano/GeoSearch/.

[3] Google Inc., *Search by Location*, 2008, http://local.google.com.

[4] Yahoo! Inc., *Yahoo local Search*, 2008, http://local.yahoo.com.

[5] Divine Inc., *Northern Light GeoSearch*. http://www.northernlight.com/geosearch.html.

[6] Overture Services, Inc., *Local Search Demo*. http://localdemo.overture.com.

[7] Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. ACM Press Book: New York, 1999.

[8] Kowalsky, G. *Information Retrieval Systems: Theory and Implementation*. Kluwer Academic Publishers: Massachusets, 1997.

[9] L. Page, S. Brin, R. Motwani e T. Winograd. *The PageRank citation ranking: Bringing order to the Web*. Technical Report SIDL-WP-1999-0120, Stanford Digital Library, 1999.

[10] Steve Lawrence e C. Lee Giles. *Searching the World Wide Web*. Science, vol.280, pp. 98-100, April 1998.

[11] W. M. Shaw Jr, R. Brugin e P. Howell. *Performance standards and evaluations in IR test collections: cluster-based retrieval models*. Information Processing & Management, 33(1):1-14, 1997.

[12] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[13] Robert Korfhage. *Information Storage and Retrieval*. John Wiley & Sons, Inc., 1997.

[14] M. Araújo, G. Navarro e N. Ziviani. *Large text searching allowing errors*. In Proceedings of WSP'97, págs 2-20, Valparaíso, Chile, 1997. Carleton University Press.

[15] S. Wartick. *Boolean Operations*. In W. B. Frakes e R. Baeza-Yates, editors, Information Retrieval: Data Structures and Algorithms, págs 264-292. Prentice Hall, Englewood Cliffs, NJ, USA, 1992.

[16] S. E. Robertson e K. Spark Jones. *Relevance weighting of search terms*. Journal of the American Society for Information Sciences, 27(3):129-146, 1976.

[17] G. Salton. *Automatic Information Organization and Retrieval*. McGraw-Hill: New York, 1968.

[18] G. Salton e M. E. Lesk. *Computer evaluation of indexing and text processing*. Journal of the ACM, 15(1):8-36, January 1968.

[19] G. Salton. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall Inc. Englewood Cliffs, NJ, 1971.

[20] L. Page, S. Brin. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. WWW7 / Computer Networks 30(1-7):107-117 (1998).

[21] Nigel Shadbolt, Wendy Hall, Tim Berners-Lee. *The Semantic Web Revisited*. IEEE Inteligent Systems 21(3) pp. 96-101.

[22] N. Guarino. *Formal Ontology and Information Systems*. in: N. Guarino, (Ed.) Formal Ontology in Information Systems. pp. 3-15, IOS Press, Amsterdam, Netherlands. 1998.

[23] Allan Heydon e Marc Najork. *Mercator: A Scalable, Extensible Web Crawler*. World Wide Web Conference, 2(4):219-229.

[24] Sriram Raghavan e Hector Garcia-Molina. *Crawling the HiddenWeb*. In Proceedings of the 27th international Conference on Very Large data Bases (VLDB), 129-138.

[25] A. Daviel. *Geotags*, April, 1999, http://geotags.com.

[26] Dublin Core Metadata Initiative. *Dublin Core Qualifiers*. Recommendation, July, 2000, http://dublincore.org/documents/dcmes-qualifiers/.

[27] K. S. McCurley. *Geospatial Mapping and Navigation of the Web*. Tenth International World Wide Web Conference, pp. 221-229. May 2001.

[28] Markowetz, Yen-Yu Chen, Torsten Suel, Xiaohui Long e Bernhard Seeger. *Design and Implementation of a Geographic Search Engine*. WebDB. Baltimore, 2005.

[29] Markowetz, Yen-Yu Chen, Torsten Suel, Xiaohui Long e Bernhard Seeger. *Design and Implementation of a Geographic Search Engine*. Technical Report TR-CIS-2005-03, CIS Department, Polytechnic University, February 2005.

[30] O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano e N. Shivakumar. *Exploiting Geographical Location Information of Web Pages*, WebDB (Informal Proceedings). 1999.

[31] N. Beckman, H.-P. Kriegel, R. Schneider e B. Seeger. *The R*-tree: An efficient and robust access method for points and retangles*. In Proceedings of SIGMOD-90, the 1990 Conference on Management of Data. 1990.

[32] Markowetz, T. Brinkhoff e Bernhard Seeger. *Exploiting the Internet As a Geospatial Database*. Workshop on Next Generation Geospatial Information. Boston, 2005.

[33] Ding, L. Gravano e N. Shivakumar. *Computing Geographical Scopes of Web Resources*. 26th International Conference on Very Large Databases, pp. 445-456. September 2000.

[34] TGN. Getty Thesaurus of Geographic Names. http://www.getty.edu/research/conducting_research/vocabularies/tgn/.

[35] Bruno Martins, Marcírio Chaves e Mário J. Silva. *Assigning geographical scopes to web pages*. ECIR, 27th European Conference on IR Research, Santiago de Compostela, Spain, March 21-23, 2005, págs 564-567.

[36] Yi Li, Alistair Moffat, Nicola Stokes e Lawrence Cavedon. *Exploring probabilistic toponym resolution for geographical information retrieval*. SIGIR, Workshop on Geographical Information Retrieval, 2006, págs 17-22.

[37] Raphael Volz, Joachim Kleb e Wolfgang Mueller. *Towards ontology-based disambiguation of geographical identifiers*. The 16th International World Wide Web Conference, Banff, Alberta, Canada, May 8-12, 2007.

[38] M. J. Silva, B. Martins, M. S. Chaves, A. P. Afonso e N. Cardoso. *Adding Geographic Scopes to Web Resources*. CEUS – Computers, Environment and Urban Systems. July, 2006.

[39] B. Martins, M. J. Silva. *A Graph-Ranking Algorithm for Geo-Referencing Documents*. 5th IEEE International Conference on Data Mining. November, 2005.

[40] Yen-Yu Chen, Torsten Suel e Markowetz. *Efficient Query Processing in Geographic Web Search Engines*. SIGMOD. Chicago, 2006.

[41] E. Amitay, N. Har'El, R. Silvan, e A. Soffer. *Web-a-where: Geotagging web content*. In Proceedings of SIGIR 2004, Workshop on Geographical Information Retrieval, págs 273–280, Sheffield, UK, June, 2004.

[42] Zhisheng Li, Chong Wang, Xing Xie, Xufa Wang e Wei-Ying Ma. *Indexing implicit locations for geographical information retrieval*. The 3rd International Workshop on Geographic Information Retrieval (GIR 2006), Seattle, USA, Aug. 2006.

[43] Chong Wang, Xing Xie, Lee Wang, Yansheng Lu, Wei-Ying Ma: *Detecting geographic locations from web resources*. The 2005 Workshop On Geographic Information Retrieval, GIR 2005, Bremen, Germany, November 4, 2005.

[44] Simon Overell, João Magalhães, Stefan Ruger. *GIR experiements with Forostar at GeoCLEF 2007*. GeoCLEF Workshop 2007, Budapest, Hungary, September 2007.

[45] Simon Overell, João Magalhães, Stefan Ruger. *Forostar: A System for GIR*. Evaluation of Multilingual and Multi-modal Information Retrieval, Berlin, Heidelberg, 2007.

[46] C. B. Jones, A.I. Abdelmoty, D. Finch, G. Fu e S. Vaid. *The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing*. In Proceedings of Third International Conference on Geographic Information Science - GIScience 2004, Maryland, USA, Lecture Notes in Computer Science 3234,125-139. 2004.

[47] S. Vaid, C. B. Jones, Hideo Joho e Mark Sanderson. *Spatio-textual Indexing for Geographical Search on the Web*. 9th International Symposium on Spatial and Temporal Databases - SSTD 2005, Lecture Notes in Computer Science 3633, 218-235. 2005.

[48] B. Martins, M. J. Silva e L. Andrade. *Indexing and Ranking in Geo-IR Systems*. Workshop on Geographic Information Retrieval at CIKM 2005. October 2005.

[49] Shashi Shekhar e Sanjay Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003.

[50] C. B. Jones, H. Alani e D. Tudhope. *Geographical information retrieval with ontologies of place*, D. Montello (ed), Spatial Information Theory Foundations of Geographic Information Science, COSIT 2001, Lecture Notes in Computer Science 2205, Springer, 323-335. 2001.

[51] Markowetz, T. Brinkhoff e Bernhard Seeger. *Geographic Information Retrieval*. 3rd Internatinal Workshop on WebDynamics. New York, 2004.

[52] M. S. Chaves, M. J. Silva e B. Martins. *A Geographic Knowledge Base for Semantic Web Applications*. Simpósio Brasileiro de Banco de Dados 2005. Uberlândia, 2005.

[53] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. *Categorizing Web Queries According to Geographical Locality*. In Proceedings of the 12th ACM Conference on Information and Knowledge Management (CIKM 2003), 2003.

[54] C. B. Jones, H. Alani e D. Tudhope. *Geographical terminology servers - closing the semantic divide*, in M.F. Goodchild, M. Duckham and M.F. Worboys (eds) Foundations of Geographic Information Science, Taylor and Francis, pp 201-218. 2003.

[55] P. D. Smart, A. I. Abdelmoty , C. B. Jones. *An Evaluation of Geo-Ontology Representation Languages for Supporting Web Retrieval of Geographic Information*. - In: Proceedings of the GIS Research UK 12th Annual Conference, Norwich, UK, pp. 175-178. 2004.

[56] C. B. Jones, A.I. Abdelmoty e G. Fu. *Maintaining ontologies for geographical information retrieval on the Web*, in Meersman, R.; Tari, Z.; Schmidt, D. C. (Eds.) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE Ontologies, Databases and Applications of Semantics, ODBASE'03, Catania, Italy, Lecture Notes in Computer Science 2888, ISBN: 3-540-20498-9 934-951. 2003.

[57] V. Gaede e O. G unther. *Multidimensional access methods*. ACM Computing Surveys, 30(2):170–231, 1998.

[58] C. Ohm, G. Klump e H. Kriegel. *Xz-ordering: A space-filling curve for objects with spatial extension*. In Proc. of the 6th Int. Symp. on Advances in Spatial Databases, págs 75–90, July 1999.

ANEXO I - Lista dos termos especiais e seus atributos

Termo	Dist. Mínima	Dist. Máxima	Tipo de Lugar	Confiança Máxima	Tipo de Referência
"cidade"	-2	-2	Município	0.800	Nome de Lugar
"município"	-2	-2	Município	0.800	Nome de Lugar
"estado"	-2	-2	Estado	0.800	Nome de Lugar
"próximo"	-2	-4		0.500	Nome de Lugar
"proximidades"	-2	-4		0.500	Nome de Lugar
"próxima"	-2	-4		0.500	Nome de Lugar
"perto"	-2	-4		0.500	Nome de Lugar
"longe"	-2	-4		0.500	Nome de Lugar
"distante"	-2	-5		0.500	Nome de Lugar
"em"	-1	-1		0.600	Nome de Lugar
"no"	-1	-1		1.000	Nome de Lugar
"subúrbio"	-2	-2	Município	0.700	Nome de Lugar
"subúrbios"	-2	-2	Município	0.700	Nome de Lugar
"periferia"	-2	-2	Município	0.700	Nome de Lugar
"centro"	-2	-2	Município	0.700	Nome de Lugar
"norte"	-2	-2		0.500	Nome de Lugar
"nordeste"	-2	-2		0.500	Nome de Lugar
"leste"	-2	-2		0.500	Nome de Lugar
"sudeste"	-2	-2		0.500	Nome de Lugar
"sul"	-2	-2		0.500	Nome de Lugar
"sudoeste"	-2	-2		0.500	Nome de Lugar
"oeste"	-2	-2		0.500	Nome de Lugar
"noroeste"	-2	-2		0.500	Nome de Lugar
"cep"	-1	-2		1.000	Código Postal
"telefone"	-1	-2		1.000	Telefone
"fone"	-1	-2		1.000	Telefone
"tel"	-1	-2		1.000	Telefone
"telefones"	-1	-10		0.800	Telefone
"fones"	-1	-10		0.800	Telefone
"tels"	-1	-10		0.800	Telefone
"endereço"	-4	-10		0.800	
"localização"	-4	-10		0.800	
"localizado"	-2	-4		0.600	
"localizada"	-2	-4		0.600	

ANEXO II - Fatores de Confiança (CF) e seus respectivos pesos, utilizados no cálculo do valor final de confiança (CR)

Local de Detecção	Tipo de Referência	Tipo de Lugar	CF	Peso
Corpo do texto	Nome de Lugar	Município e Estado	CF_{CASE}	10%
			CF_{ABV}	10%
			CF_{ST}	25%
			CF_{TS}	25%
			CF_{CROSS}	30%
		Outros	Constante	20%
			CF_{ST}	25%
			CF_{TS}	25%
			CF_{CROSS}	30%
	Sigla de Unidade Federativa (UF)	Estado	Constante	30%
			CF_{CROSS}	70%
	Código Postal	Município	CF_{FMT}	20%
			CF_{ST}	60%
			CF_{CROSS}	20%
	Código de Área Telefônico	Estado	CF_{FMT}	20%
			CF_{ST}	60%
			CF_{CROSS}	20%
	Gentílico	Município e Estado	Constante	40%
			CF_{CROSS}	60%
Título da Página	Nome de Lugar	Município e Estado	CF_{CASE}	10%
			CF_{ABV}	10%
			CF_{TS}	40%
			CF_{CROSS}	40%
		Outros	CF_{TS}	60%
			CF_{CROSS}	40%
	Gentílico	Município e Estado	Constante	40%
			CF_{CROSS}	60%
URL	Nome de Lugar	Todos	CF_{TS}	60%
			CF_{CROSS}	40%
	Gentílico	Município e Estado	Constante	40%
			CF_{CROSS}	60%

ANEXO III - Referências Geográficas citadas e seus respectivos valores para o fator de confiança CF_{TS}

Referência	Tipo de Lugar	CF_{TS}
Qualquer referência	Estado	1,000
Aliança, Antônio Cardoso, Capim, Centenário, Dionísio, Esmeralda, Peixe, Prata, Quatro Irmãos, Telha, Travesseiro	Município	0,000
Aracaju	Município	0,614
Aracaju	Microrregião	0,105
Barro	Município	0,137
Barro	Microrregião	0,158
Bela Vista, Olinda	Município	0,204
Conquista	Município	0,614
Descanso, Marco, Ouro	Município	0,068
Leste Sergipano	Mesorregião	0,882
Litoral Norte	Microrregião	0,158
Noroeste Rio-Grandense	Mesorregião	0,471
Oeste Catarinense	Mesorregião	0,412
Própria	Município	0,068
Própria	Microrregião	0,263
Recife	Município	0,750
Recife	Microrregião	0,211
Santos	Município	0,204
Santos	Microrregião	0,158
São Pedro	Município	0,683
Uberlândia	Município	0,341
Uberlândia	Microrregião	0,158
Valença	Município	0,546
Valença	Microrregião	0,105

ANEXO IV - Localidades referenciadas e número de ocorrências por documento.

Localidade	Tipo	Qtd	Localidade	Tipo	Qtd
Alagoas	4	4	Minas Gerais	4	24
Acre	4	8	Monte Alto (SP)	1	2
Altamira (PA)	1	8	Monte Azul Paulista (SP)	1	2
Alta Floresta (MT)	1	2	Norte de Minas (MG)	3	2
Amapá	4	4	Nova Lacerda (MT)	1	2
Amazonas	4	14	Olímpia (SP)	1	2
Araçatuba	1	2	Pará	4	32
Araguaína (TO)	1	2	Paraíba	4	2
Arguianópolis (TO)	1	2	Paraná	4	16
Bahia	4	8	Paranaíba (MS)	1	2
Belém (PA)	1	4	Paulista (PE)	1	8
Belo Horizonte (MG)	1	8	Paulistana (PI)	1	4
Brasil Novo (PA)	1	2	Pernambuco	4	6
Brasília (DF)	1	2	Piauí	4	4
Breu Branco (PA)	1	2	Piranga (MG)	1	2
Cajobi (SP)	1	2	Porto Alegre (RS)	1	56
Campo Grande (MS)	1	2	Porto Firme (MG)	1	2
Carmo (RJ)	1	2	Recife (PE)	1	2
Ceará	4	4	Resplendor (MG)	1	2
Comodoro (MT)	1	2	Ribeirão (PE)	1	2
Cuiabá (MT)	1	6	Ribeirão Preto (SP)	1	4
Diadema (SP)	1	2	Rio de Janeiro	4	20
Distrito Federal	4	4	Rio de Janeiro (RJ)	1	20
Dourados (MS)	1	2	Rio Grande do Sul	4	8
Duque de Caxias	1	2	Rondônia	4	34
Embu (SP)	1	22	Roraima	4	2
Espírito Santo	4	2	Roraima	4	2
Estreito (MA)	1	2	Santa Catarina	4	6
Fernando de Noronha (PE)	1	2	Santana do Riacho (MG)	1	2
Fernando de Noronha (PE)	2	2	São Bernardo do Campo (SP)	1	2
Fortaleza (CE)	1	2	São José do Rio Preto (SP)	1	2
Goiânia (GO)	1	2	São Paulo	4	96
Goiás	4	6	São Paulo (SP)	1	94
Itacoatiara (AM)	1	2	Sapezal (MT)	1	2
Itaguaí (RJ)	1	2	Socorro (SP)	1	2
Joinville (SC)	1	2	Sul Espírito-Santense	3	4
Juruena (MT)	1	2	Tailândia (PA)	1	2
Manaus (AM)	1	2	Tocantins	4	10
Magaratiba (RJ)	1	2	Três lagoas (MS)	1	2
Mairinque (SP)	1	2	Tucuruí (PA)	1	2
Manicoré (AM)	1	2	Uberlândia (MG)	1	2
Maranhão	4	14	Vila Bela da Santíssima Trindade (MT)	1	2
Marília (SP)	3	2	Vitória (ES)	1	4
Mato Grosso	4	36	Votuporanga (SP)	1	2
Mato Grosso do Sul	4	12	Zona da Mata (MG)	3	2

Tipo: 1=Município; 2=Microrregião; 3=Mesorregião; 4=Estado; 5=Região.