Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

UM *GAZETTEER* ONTOLÓGICO PARA RECUPERAÇÃO DE INFORMAÇÃO GEOGRÁFICA

Ivre Marjorie Ribeiro Machado

Belo Horizonte
2011

UM *GAZETTEER* ONTOLÓGICO PARA RECUPERAÇÃO DE INFORMAÇÃO GEOGRÁFICA

IVRE MARJORIE RIBEIRO MACHADO

UM *GAZETTEER* ONTOLÓGICO PARA RECUPERAÇÃO DE INFORMAÇÃO GEOGRÁFICA

Dissertação apresentada ao Departamento de Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: CLODOVEU AUGUSTO DAVIS JÚNIOR

Belo Horizonte

2011

Machado, Ivre Marjorie Ribeiro.

M149g Um *gazetteer* ontológico para recuperação de informação geográfica / Ivre Marjorie Ribeiro Machado. — Belo Horizonte, 2011.
xxi, 85f. : il.; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais. Departamento de Ciência da Computação.

Orientador: Clodoveu Augusto Davis Júnior.

1. Computação - Teses. 2. Sistemas de recuperação da informação - Teses. 3. Sistemas de informação geografica – Teses. I. Orientador. II. Título.

CDU 519.6*74 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Um gazetteer ontológico para recuperação de informação geográfica

IVRE MARJORIE RIBEIRO MACHADO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROF. RICARDO DA SILVA TORRES
Instituto de Computação - UNICAMP

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 13 de maio de 2011.

Dedico este trabalho a minha querida mãe, que sozinha me fez ser quem eu sou.

Agradecimentos

Agradeço a Deus por me guiar e não permitir que eu desistisse nos momentos de fraqueza.

Ao meu orientador, Clodoveu Davis pelas sabias palavras, orientações, tranqüilidade e serenidade que conseguiu me passar durante toda a realização desse trabalho.

Ao meu marido, companheiro para todas as horas, por mais uma vez ter tanta paciência e entender que os momentos de nervosismos iriam passar e tudo voltaria ao normal. Além disso, por ter me ajudado no desenvolvimento desse trabalho.

À minha mãe por entender a minha ausência e distância durante esses dois anos de mestrado e sempre me apoiar nas minhas decisões.

À minha família pela minha grande ausência nas reuniões e aniversários, além das rápidas visitas. Que essa ausência seja um sinônimo de alegria para vocês no futuro. Adoro vocês!

Aos meus amigos do LBD (Laboratório de Banco de dados), em especial ao Leo, Hugo, Daniel Hasan e Rafael Odon, por me ajudarem tanto e pelos momentos que passamos juntos, mesmo que cada um no seu cantinho.

Às minhas amigas, Glívia, Polly, Letícia, Elisa por agüentarem tantas angustias, reclamações e medos. E a todos outros amigos pela minha ausência e distância.

Aos funcionários do DCC, em especial Renata, Sheila, Túlia e Maristela pelo excelente e organizado trabalho.

Enfim, a todos que de alguma forma contribuíram para o meu crescimento e para a conclusão desse importante trabalho de mestrado!

Resumo

O volume de informações espaciais na Web cresce a cada dia, tanto na forma de mapas, como de referências a lugares em documentos e páginas. Considerando as necessidades de informação espaciais dos usuários, muitas vezes é necessário determinar a que lugares o texto da página se refere. Este trabalho apresenta uma proposta de uma nova geração de *gazetteer* (dicionário toponímico), que inclui elementos tais como relacionamentos espaciais, conceitos e termos relacionados, formando essencialmente uma ontologia de lugares. Esse *gazetteer* ontológico, denominado OntoGazetteer, fornece suporte semântico para resolver vários problemas comuns na recuperação de informação geográfica. A validação da estrutura proposta para o OntoGazetteer foi feita a partir de estudos de caso que abordaram duas categorias de problemas de recuperação de informação geográfica: (1) detecção e inferência do contexto geográfico em textos e (2) desambiguação de nomes de lugares. Esses estudos mostraram bons resultados, recomendando o uso do OntoGazetteer em problemas de recuperação de informação geográfica.

Palavras chave: *Gazetteer*, Recuperação de Informação Geográfica, Ontologias Espaciais, Reconhecimento de Nomes de Lugares, Desambiguação de Nomes de Lugares.

Abstract

The volume of spatial information on the Web grows continuously, both as maps and as references to places in documents and pages. Considering user needs for spatial information, it is often necessary to determine the places to which the text in a page refers. This work introduces a proposal for a new generation of gazetteers (place names dictionaries) that includes elements such as spatial relationships, concepts and related terms, essentially forming an ontology of places. This ontological gazetteer, or OntoGazetteer, provides semantic support for solving various common problems in geographic information retrieval. The validation of the proposed structure for the OntoGazetteer was achieved through case studies that covered two categories of geographic information retrieval problems: (1) detection and inference of the geographic context in texts, and (2) disambiguation of place names. These studies showed good results, which recommend the use of the OntoGazetteer in further geographic information retrieval problems.

***Keywords**: Gazetteer, Geographic Information Retrieval, Spatial Ontologies, Recognition of place names, Place Name Disambiguation*

Lista de figuras

Lista de tabelas

Lista de algoritmos

Lista de siglas

ADL - Alexandria Digital Library

CASNAV - Centro de Análises de Sistemas Navais

GKB - Geographic Knowledge Base

GIR - Geographic Information Retrieval

IBGE - Instituto Brasileiro de Geografia e Estatística

OWL - Web Ontology Language

RDF - Resource Description Framework

RI - Recuperação de Informação

TGN - Getty Thesaurus of Geographic Names

XML - Extensible Markup Language

Sumário

Capítulo 1

Introdução

1.1 Contextualização do Problema

Atualmente, o volume de informação disponível é muito grande e cresce todos os dias. Um exemplo disso é a informação na Web. A recuperação dessas informações exige sistemas que sejam capazes de compreender as necessidades dos usuários, localizar os documentos relevantes e apresentar para os usuários esses documentos através de um *ranking* de relevância. Este é um tipo de tarefa associada aos sistemas de recuperação de informação, que também tratam de questões relativas à indexação e armazenamento de documentos.

Os usuários manifestam suas necessidades de recuperação de informação de muitas maneiras, mas principalmente através de palavras-chave submetidas a uma máquina de busca. Trabalhos anteriores [Backstrom, Kleinberg, Kumar, & Novak, 2008; Delboni, Borges, Laender, & Davis-Jr, 2007; Sanderson & Kohler, 2004; L. Wang et al., 2005] mostraram que uma parcela significativa dessas consultas envolvem termos ou expressões com significado espacial, incluindo nomes de lugares e expressões de linguagem natural que indicam posicionamento. No entanto, obter resultados significativos de tais consultas é muitas vezes difícil, porque as palavras-chave geograficamente relevantes às vezes não são entendidas como tal pelos sistemas de recuperação de informação. Técnicas de recuperação de informação geográfica têm limitações importantes no reconhecimento das referências espaciais e em lidar com

nomes ambíguos (por exemplo, "São Paulo" pode ser um estado brasileiro, uma cidade, ou um time de futebol). Existem também algumas dificuldades na recuperação de informações restritas a um contexto geográfico. Se o documento puder ser associado a um conjunto de lugares, é possível ajustar a sua posição em um *ranking*, ou filtrar documentos que se referem a locais indesejados. O reconhecimento de um termo como sendo o nome de um lugar é em geral realizado com o apoio de um *gazetteer*, um dicionário de nomes de lugares [Hill, 2000].

Atualmente, alguns *gazetteers* estão disponíveis na Web. No entanto eles possuem estruturas de dados muito simples, com apenas três componentes: o nome do lugar, o seu tipo (definido como membro de uma taxonomia de tipos), e o *footprint* (em geral, um par de coordenadas que indicam a sua localização). Com este tipo de estrutura, os *gazetteers* apresentam várias limitações, que tornam difícil o uso deles em problemas de recuperação de informação. Além disso, o conteúdo dos *gazetteers* em geral não inclui nomes de lugares intra-urbanos, tais como nomes de ruas, de bairros, pontos de referência e atrações turísticas, e não existem recursos que permitam registrar e utilizar o relacionamento espacial entre seus elementos, com exceção de estimativas de proximidade baseadas nas coordenadas.

Gazetteers são difíceis de manter e expandir, e com isso seu conteúdo, geralmente, é irregular: embora alguns incluam detalhes urbanos de cidades dos EUA ou Europa, lugares brasileiros não estão bem cobertos. Algumas dessas dificuldades podem ser superadas pela utilização de serviços de geocodificação, como os disponíveis na API do Google Maps, que não fazem entradas explícitas em *gazetteers*, mas são capazes de fornecer um ou mais pares de coordenadas que correspondem a uma descrição textual.

1.2 Justificativa

Mesmo com a estrutura limitada, diversas aplicações geográficas baseadas na Web utilizam os *gazetteers*, como mostram Goodchild e Hill [2008]. Isso mostra a necessidade e importância em desenvolver estudos nessa área e fazer evoluir a estrutura desse mecanismo. Jones e Purves [2008] apresentam áreas e problemas de pesquisa na área de recuperação de informação geográfica, e multiplicam seu crescimento pelo fato

de que grande parte da informação disponível na Web tem ligação com localizações geográficas. Um mecanismo capaz de auxiliar na solução desses problemas, ou seja, na identificação de termos como nomes de lugares, o tipo do lugar e a coordenada geográfica ao qual se refere, são os tradicionais *gazetteers*.

1.3 Objetivos

O objetivo principal deste trabalho consiste em propor uma nova estrutura para *gazetteers*, e implementá-la sob a forma de uma ontologia de lugares, capaz de apoiar o desenvolvimento de projetos de recuperação de informação geográfica. Um *gazetteer ontológico*, como será demonstrado, permite não apenas identificar nomes de lugares, mas também registrar conceitos e termos relacionados a um lugar, como em uma ontologia em que os conceitos principais são os lugares e suas características. Assim o *gazetteer* proposto tem, como principais elementos, lugares e seus relacionamentos, em analogia aos conceitos ou classes e suas propriedades e relacionamentos registrados em ontologias.

Para alcançar o objetivo geral são necessários os seguintes objetivos específicos:

- Propor uma nova estrutura para os *gazetteers*;
- Implementar a estrutura proposta sob a forma de uma ontologia de lugares;
- Carregar dados usando a nova estrutura (neste trabalho nos limitamos a dados sobre o Brasil);
- Avaliar o *gazetteer* Ontológico através de estudos de caso com aplicações de recuperação de informação geográfica;
- Criar uma interface e serviços Web que facilitem o acesso aos dados do *gazetteer* Ontológico para futuras atividades e aplicações.

1.4 Motivação

O *gazetteer* ontológico proposto foi concebido para apoiar a solução de problemas de recuperação de informação geográfica, tais como reconhecimento do contexto geográfico de textos a partir da semântica associada aos lugares e dos relacionamentos

entre eles, desambiguação de nomes de lugares, *ranking* de relevância de documentos com o foco geográfico, desenvolvimento de métodos para avaliar sistemas de recuperação de informação geográfica, entre outros apresentados ao longo do trabalho [Adriani & Paramita, 2007; S. E. Overell & Ruger, 2007; Silva, Martins, Chaves, Afonso, & Cardoso, 2004; Wang, Xie, Wang, Lu, & Ma, 2005]. O presente trabalho aborda a concepção e construção do *gazetteer*, porém apenas o aplica experimentalmente na solução de um subconjunto desses problemas.

1.5 Metodologia

A metodologia utilizada neste trabalho foi inicialmente de natureza exploratória, através da revisão da literatura, com o objetivo de compreender melhor a estrutura e limitações dos *gazetteers* tradicionais.

Após realizar a revisão bibliográfica, um esquema conceitual com uma nova estrutura foi proposta para os *gazetteers*. Esse esquema foi implementado usando um sistema de gerenciamento de banco de dados geográficos. Em seguida, dados geográficos de diversas fontes foram carregados, de forma em geral manual, para possibilitar a validação e uso do *gazetteer* ontológico.

A avaliação do OntoGazetteer assim criado foi feita através de dois estudos de caso, no qual foram escolhidas duas tarefas em recuperação de informação geográfica para serem implementadas com o apoio do *gazetteer* ontológico. Alguns experimentos foram feitos e verificados com o apoio de voluntários. Por fim, os resultados foram computados e avaliados.

1.6 Organização deste documento

O restante desse trabalho está organizado da seguinte maneira.

No Capítulo 2 é apresentado o levantamento bibliográfico sobre os temas relacionados a esse trabalho, além dos trabalhos mais relevantes e que foram referenciais para o desenvolvimento deste. Além de apresentar possíveis aplicações em problemas de recuperação de informação geográfica que podem ser apoiadas por um

gazetteer.

No Capítulo 3 é descrita a estrutura proposta para o *gazetteer* sob a forma de ontologia de lugar, bem como sua implementação. Além disso, descreve como o *gazetteer* ontológico ou OntoGazetteer pode apoiar tarefas de recuperação de informação geográfica. A interface Web para acesso aos dados do *gazetteer* ontológico é apresentada.

No Capítulo 4 são apresentados dois estudos de caso para avaliar a estrutura proposta para o OntoGazetteer. Os estudos de caso abordaram duas categorias de problemas de recuperação de informação geográfica: (1) detecção e inferência do contexto geográfico em textos e (2) desambiguação de nomes de lugares. Além de descrever os resultados dos experimentos.

Finalmente, no Capítulo 5 é realizado um registro dos benefícios alcançados e dos trabalhos futuros propostos.

Capítulo 2

Referencial Teórico

Esse capítulo apresenta uma revisão dos principais conceitos relacionados a este trabalho. Além disso, descreve publicações com contribuições relevantes e que influenciaram o desenvolvimento do trabalho.

2.1 Recuperação de Informação Geográfica

A recuperação de informação (RI) ou "*information retrieval*" lida com a representação, o armazenamento, a organização e o acesso aos itens de informação. A representação e organização dos itens de informação devem prover ao usuário um fácil acesso a informação o qual ele está interessado. No entanto, a caracterização da informação que o usuário necessita não é um problema simples. Sendo assim, é necessário oferecer ao usuário uma interface que o permita traduzir a sua necessidade por informação na forma de uma consulta (*query*) que poderá ser processada pela máquina de busca ou sistema de recuperação de informação [Baeza–Yates & Ribeiro-Neto, 1999].

Segundo Baeza–Yates e Ribeiro-Neto [1999], a recuperação de dados no contexto de um sistema de RI consiste em determinar os documentos que contêm determinadas palavras chave na consulta feita pelo o usuário. Isso não é suficiente para atender a necessidade do usuário por informação. Sendo assim, os sistemas de RI se concentram em recuperar a informação relativa a um determinado assunto de uma consulta e não apenas os dados. Por isso, muitos algoritmos e técnicas são criados com o objetivo de

compreender cada vez melhor a necessidade do usuário e retornar o melhor resultado. Essas técnicas incluem recursos para desambiguação, classificação, filtragem, expansão de consultas, *ranking*, entre outros. Em resumo, a desambiguação é usada para distinguir entre significados diferentes de um termo; a classificação procura agrupar informações correlacionadas; a filtragem busca selecionar dados de acordo com um critério definido; a expansão de consulta acrescenta informação a uma consulta para tentar obter um número maior de resultados relevantes; e o *ranking* é um processo de ordenação dos resultados de uma consulta para que os mais relevantes fiquem em destaque.

Como, nas últimas décadas, o volume de documentos e de informação tem crescido muito, principalmente na Web, a recuperação eficiente de informação ainda é um problema relevante e, consequentemente, constitui uma importante área de pesquisa. Um motor de busca ou máquina de busca é um sistema de RI voltado para a Web que possui uma interface que possibilita ao usuário realizar uma consulta, que será traduzida em busca de entender sua necessidade por informação, fornecendo como resultado um conjunto ordenado de páginas e documentos [Baeza–Yates & Ribeiro-Neto, 1999]. Atualmente existem muitas máquinas de busca que implementam os algoritmos e técnicas descritos. Alguns exemplos são Google[1] e Yahoo[2].

Em uma parcela significativa das consultas realizadas a máquinas de busca, usuários pretendem obter informação sobre lugares ou correlacionar assuntos a lugares específicos. No entanto, as máquinas de busca apenas oferecem a possibilidade de se informar palavras-chave, e portanto resultados inadequados podem ser obtidos quando termos são ambíguos, ou seja, o lugar procurado tem nome coincidente com algum objeto ou pessoa, e isso ocorre com freqüência. A Figura 2-1 apresenta parte do resultado de uma consulta por “São Paulo” na máquina de busca Google. Como não é possível informar a busca apenas por lugares, resultados relacionados a time de futebol também são retornados.

[1] www.google.com.br/

[2] www.yahoo.com/

Figura 2-1 - Resultado de consulta no Google (fonte: Google, 15/4/2011)

Estudos realizados mostram que é grande o número de palavras referentes à geografia entre as palavras-chaves apresentadas a diversas máquinas de busca. Sanderson e Kohler [2004] realizaram um estudo com um *log* da máquina de busca Excite, no qual foi verificado que 19% de 2.500 consultas possuíam pelo menos um termo geográfico. Outro trabalho realizado sobre a máquina de busca brasileira TodoBR verificou que 14% das consultas de um *log* de 6 meses tinham pelo menos um termo relacionado a geografia [Delboni et al., 2007].

Para responder de forma adequada às consultas que demandam informação geográfica, seria necessário aperfeiçoar as máquinas de busca em dois aspectos principais. O primeiro seria passar a reconhecer a intenção geográfica por trás de um conjunto de palavras-chave, quando estas incluem termos geográficos. O segundo seria incluir informação de localização, obtida a partir do conteúdo das páginas, aos índices, o que implica a necessidade de reconhecer o contexto geográfico de documentos da Web, quando este existe. Intuitivamente, percebemos que boa parte dos documentos de interesse para os sistemas de RI possui referências geográficas textuais. Alguns

exemplos são as notícias de jornais, páginas e outros documentos da Web. Notícias, de modo geral, como parte da técnica de redação empregada por jornalistas, apresentam o nome do lugar no qual ocorreu o fato descrito. No caso de páginas da Web, em geral, muitas vezes se pode inferir a localização a partir de elementos tais como códigos postais ou códigos de área de telefonia, ou mesmo a partir da ocorrência de endereços postais completos [Borges, 2006; Borges, Laender, Medeiros, & Davis Jr, 2007]. Em outros casos, a ocorrência de termos típicos de determinados lugares permitem inferir a correlação do texto com determinado lugar [Alencar, Davis-Jr., & Gonçalves, 2010]. Por exemplo, uma referência a um prato típico ou a uma manifestação cultural pode conduzir o leitor a inferir um lugar associado. Em casos menos comuns, a referência à localização pode ser explicita, como no caso de *tags* contendo coordenadas geográficas associadas a fotos ou vídeos em sites especializados.

Esse é o universo de trabalho da recuperação de informação geográfica. A recuperação de informação geográfica lida com aspectos de duas áreas, a recuperação de informação e os sistemas de informação geográfica [Brisaboa, Luaces, Places, & Seco, 2010; Jones & Purves, 2008]. Segundo Brisaboa, Luaces et al. [2010], a recuperação de informação geográfica requer muita pesquisa, já que muitos dos algoritmos e técnicas utilizados em RI não levam em consideração aspectos geográficos.

No entanto, na recuperação de informação geográfica, como em qualquer recuperação de informação, o usuário deseja uma resposta com informação relevante. Isso significa que uma busca pela cidade de São Paulo, que resulte em um documento sobre o São Paulo Futebol Clube, provavelmente não atenderá às expectativas do usuário (ver Figura 2-1). Dessa forma, é necessário o uso de mecanismos e métodos que auxiliem na identificação de entidades geográficas em diversas situações: entre palavras-chave de uma consulta, em textos, em *tags* para que a recuperação de informação geográfica seja mais eficiente. Segundo Janowicz e Kessler [2008], um mecanismo muito utilizado para auxiliar na recuperação de informação geográfica é o *gazetteer* digital, que é um dicionário de nomes de lugares.

2.2 *Gazetteer* Digital

Gazetteer digital ou simplesmente *gazetteer*, é um dicionário geoespacial de nomes geográficos. Também conhecido como dicionário toponímico ou dicionário de lugares, é uma versão digital dos índices de topônimos, geralmente encontrados em Atlas.

Em geral, *gazetteers* possuem uma estrutura simples, com apenas três componentes: o nome do lugar, a localização e o tipo do lugar, como mostra a Figura 2-2. O nome do lugar identifica lugares a partir de uma descrição textual. Podem ser incluídas variações comuns, tais como abreviaturas e apelidos, e também nomes antigos ou anteriores, e nomes em outras línguas. Já a localização, também chamada *footprint*, é o componente capaz de associar determinado nome a uma coordenada geográfica. Em geral, os *gazetteers* trazem apenas um par de coordenadas associadas a cada lugar catalogado, e não uma delimitação geométrica completa. O tipo do lugar é uma classificação responsável por diminuir problemas de ambigüidade, devido ao fato de um mesmo nome identificar diferentes lugares de naturezas diferentes [Hill, 2000].

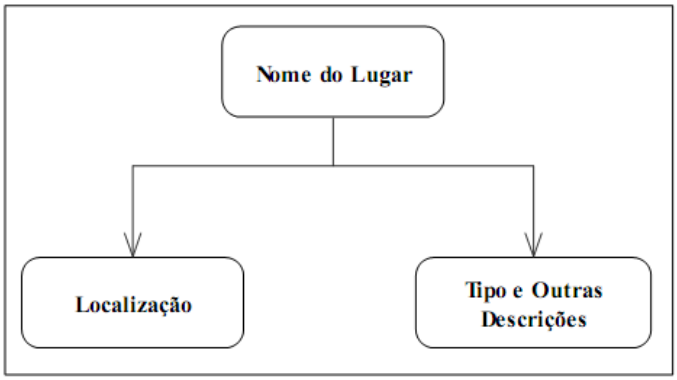


Figura 2-2 - Componentes básicos do *gazetteer* (fonte: [Gouvêa, 2009])

O tipo de lugar é parte de uma taxonomia ou hierarquia pré-definida, que varia entre os *gazetteers*. Por exemplo, o nível mais alto da hierarquia de tipos definidos pelo *gazetteer* da ADL[3] (*Alexandria Digital Library*) inclui áreas administrativas, recursos

[3] http://www.alexandria.ucsb.edu/

hídricos, elementos de relevo e físicos. Esses níveis podem ser especializados em mais três níveis. Por outro lado, o *gazetteer* GeoNames[4] define sua própria hierarquia de tipos, cujo primeiro nível possui nove classes, com apenas um nível de especialização subsequente. Existe uma proposta para criar um dicionário padrão de feições para o território brasileiro [Barbosa & Casanova, 2010].

Alguns *gazetteers* disponibilizam suas bases na Internet, tais como GeoNames, ADL *Gazetteer*, TGN[5] (*Getty Thesaurus of Geographic Names*) e GKB[6] (*Geographic Knowledge Base*). A Tabela 2-1 apresenta uma breve comparação entre esses dicionários de nomes de lugares.

Tabela 2-1 - Comparação entre *gazetteers*

Gazetteer	**Qtde Nomes**	**Representação**	**Informação**
GeoNames	Mais de 8 milhões de nomes	9 classes e 635 subclasses	Possui nomes de lugares alternativos e usa fontes de dados públicas dos Estados Unidos e a Wikipédia.
ADL *Gazetteer*	5,9 milhões de nomes	Diversas categorias	A ADL é uma biblioteca digital com diversas coleções de dados, entre elas, um *gazetteer.*
TGN	Cerca de 1 milhão de nomes	Hierárquico	Muito utilizado em vários estudos e sistemas como é o caso da máquina de busca SPIRIT [Jones et al., 2002].
GKB	Cerca de 34 mil nomes	Diversas categorias	Desenvolvido no projeto GREASE da Universidade de Lisboa, contém apenas termos geográficos de Portugal e duas ontologias: Geo-Net-PT01 e PT02 [Chaves, 2009; Lopez-Pellicer, Silva, & Chaves, 2010]

Diversas tarefas são realizadas com o apoio dos *gazetteers*, tais como reconhecimento de entidades, desambiguação de nomes, atribuição de nomes, classificação de documentos, entre outras. Amitay, Har'El et al. [2004] apresentam o Web-a-Where, um sistema que identifica *tags* geográficas para páginas Web, com o auxílio de um *gazetteer*. Souza, Davis-Jr et al. [2005] e Souza, Delboni et al. [2004]

[4] http://www.geonames.org/

[5] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

[6] http://xldb.fc.ul.pt/wiki/Grease

apresentam o Locus, um sistema de localização geográfica, que possui um *gazetteer* como um componente central e que foi baseado em uma ontologia, a OnLocus [Borges, 2006; Borges et al., 2007]. Buscaldi e Rosso [2007] apresentam uma comparação entre dois métodos para identificação automática de localizações em artigos da Wikipédia. Overell e Ruger [2008] descrevem um modelo baseado na co-ocorrência para resolver o problema de desambiguação de nomes de lugares. Para isso, usam uma combinação de *gazetteers* e heurísticas.

Outros trabalhos apresentam propostas para construção automática de *gazetteers*, já que a carga e a manutenção manual dessas bases de dados geram muito trabalho, além de se tornarem progressivamente mais difíceis, ao longo do tempo, com o crescimento da base armazenada [Popescu, Grefenstette, & Moellic, 2008; Toral & Munoz, 2006; Uryupina, 2003]. Estes trabalhos extraem dados da Wikipédia, que é um grande repositório de informação disponível em diversos idiomas. Gouvêa, Loh et al. [2008] apresentam uma estratégia para identificar entidades encontradas em notícias para serem usadas no desenvolvimento de *gazetteers*. Alencar, Davis-Jr. et al. [2010] descrevem uma estratégia para classificação de texto que utiliza dados extraídos da Wikipédia que podem ser consideradas evidências de nomes de lugares.

Gazetteers são recursos importantes para a implementação de algoritmos de geocodificação, que é o processo de transformação de descrições textuais em coordenadas. São também importantes em processos de geocodificação reversa, que consiste em obter uma descrição de um lugar a partir de coordenadas dadas [Janowicz & Kessler, 2008]. Goodchild e Hill [2008] mostram que o estudo de dicionários toponímicos é ainda um tema muito importante e que tem gerado muita pesquisa. Os autores também salientam, como principais itens de estudo dos *gazetteers*, os seus componentes, o processo de nomear um lugar e a interoperabilidade entre eles.

Apesar de serem um importante mecanismo para auxiliar na recuperação de informação geográfica, os *gazetteers* apresentam limitações, sendo as principais [Borges, 2006; Borges et al., 2007; Fu, Jones, & Abdelmoty, 2005; Souza et al., 2005]:

- Possuem limitações conceituais que reduzem sua utilidade para recuperação de informação;
- Restringem a representação espacial dos objetos referenciados a uma geometria simples (*footprint* – um ponto ou um retângulo);

- Não possuem suporte para localizações semanticamente completas, mas geograficamente imprecisas, tais como “centro de Belo Horizonte” ou “Sul do Brasil”;
- Não possuem nomes de lugares intraurbanos, tais como nomes de ruas, de bairros, pontos de referência atrações turísticas, entre outros.

Além disso, os *gazetteers* disponíveis na Web não contemplam de forma ampla lugares geográficos (municípios, bairros, ruas) de países específicos como é o caso do Brasil [Gouvêa, 2009; Gouvêa et al., 2008]. Diante dessas limitações e da importância do uso desse mecanismo, fica clara a necessidade da criação de uma nova geração de *gazetteers* que permita sua efetiva utilização em recuperação de informação geográfica.

2.3 Problemas de Recuperação de Informação Geográfica

Devido ao crescimento explosivo da Internet, a recuperação de informação (RI) tem sido o foco mais recente de pesquisas. Como grande parte da informação disponível possui informação geográfica, técnicas e algoritmos de RI são expandidos para tratar problemas tais como detecção de referências a lugares ou associação de localizações a documentos Web, portanto aplicações com foco em recuperação de informação geográfica. Alguns desses problemas foram destacados por [Jones & Purves, 2008] em uma pesquisa em *geographic information retrieval* (GIR):

- Detecção de referências geográficas na forma de nomes de lugares;
- Desambiguação de nomes de lugares;
- Interpretação geográfica de nomes de lugares vagos, tais como “Sul da França”;
- Indexação de documentos de acordo com o contexto geográfico e de conteúdo não-espaciais;
- *Ranking* de documentos por relevância geográfica;
- Melhoria da interface de busca;
- Avaliação de métodos para comparar de sistemas e técnicas de GIR.

Os *gazetteers* podem ser usados como mecanismos para fornecer um suporte ou solucionar muitos desses problemas em recuperação de informação geográfica.

Entretanto, os tradicionais *gazetteers* possuem limitações que podem reduzir o seu uso ou até mesmo não fornecer dados suficientes para auxiliar na solução desses problemas. A seguir são descritos alguns dos problemas de GIR e como os *gazetteers* podem ser usados.

2.3.1 Detecção de referências geográficas

Geoparsing é o processo de analisar um texto e identificar referências para lugares na forma de nomes de lugares e outros termos relacionados com o espaço geográfico [Jones & Purves, 2008]. Por outro lado, *geotagging* é o processo de identificação de entidades geográficas mencionadas direta e indiretamente em um texto e a criação de *tags* que permitem a vinculação de documentos a um lugar ou conjuntos de lugares [Amitay et al., 2004; Teitler et al., 2008]. Ambos *geoparsing* e *geotagging* requerem o reconhecimento de referências geográficas em texto, e se essa tarefa é feita corretamente, o contexto geográfico do documento pode ser facilmente estabelecido.

Os *gazetteers* mantêm listas de nomes de lugares oficiais e alternativos que facilitam na identificação de nomes de lugares candidatos encontrados em textos. No entanto, isso não é o bastante para realizar *geoparsing* e *geotagging*. Alguns *gazetteers* até mantêm informações sobre as hierarquias espaciais e o *footprint* permite a identificação de lugares adjacentes. No entanto, para esse tipo de identificação é necessário o uso de funções geográficas o que pode necessitar de um grande processamento que pode ser bastante lento dependendo da base analisada.

Uma das limitações dos tradicionais *gazetteers* está relacionada ao fato de eles não manterem os relacionamentos entre os lugares em suas bases. Isso pode dificultar a realização de algumas inferências como, por exemplo, um documento que menciona várias cidades em um mesmo estado, tem como contexto real o próprio estado.

2.3.2 Desambiguação de nomes de lugares

O processo de mapear entidades para uma representação espacial (coordenadas

geográficas) dado um contexto é conhecido como resolução de topônimos e, segundo Leidner [2004], é um pré-requisito para realizar a recuperação de informação geográfica com alta qualidade. A inferência de contexto geográfico em textos necessita de um reconhecimento correto de nomes de lugares (topônimos). Sendo assim, é necessário identificar termos que sejam possíveis referências a lugares, e distingui-los de termos iguais com semântica diferente, ou seja, realizar sua desambiguação em relação a outros lugares que tenham o mesmo nome e em relação a outras entidades com nome coincidente.

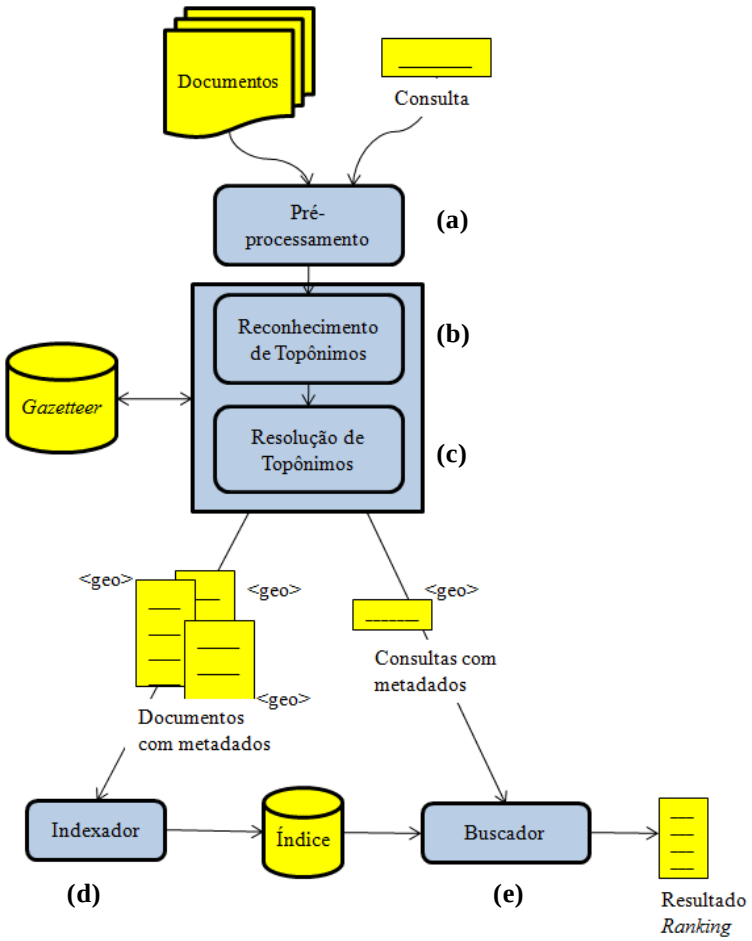


Figura 2-3 - Componentes de um GIR (fonte: Adaptação de [Leidner, 2007])

Um sistema de recuperação de informação geográfica como mostra a Figura 2-3 possui alguns componentes principais, além de uma sequência em todo o seu processo,

que começa com pré-processamento de documentos e consultas até a apresentação do resultado através de um *ranking*. O primeiro componente (Figura 2-3 (a)) é responsável por realizar o pré-processamento dos documentos ou consultas. Esse pré-processamento é responsável por excluir *stopwords*, que são palavras comuns como artigos, preposições e conjunções, portanto, palavras que podem ser consideradas irrelevantes para o contexto do sistema de recuperação de informação geográfico. O segundo componente (Figura 2-3 (b)) é responsável por reconhecer possíveis termos geográficos com auxilio de um *gazetteer*. O terceiro componente (Figura 2-3 (c)), o responsável pela resolução de topônimos, usa estratégias para resolver ambigüidades entre esses termos anteriormente reconhecidos como possíveis candidatos a nomes de lugares. O próximo componente (Figura 2-3 (d)) é o indexador, que é responsável por criar índices que melhorem a qualidade das buscas realizadas no sistema. Por fim, o último componente (Figura 2-3 (e)) é responsável por realizar a busca nos índices e apresentar o resultado na forma de um *ranking* de relevância.

Como os nomes de lugares são frequentemente ambíguos, a ambigüidade é um importante problema na recuperação de informação geográfica, que deve ser tratado. O nome “São Paulo” possui 6.522 registros no Geonames. Segundo Smith e Crane [2001], o *gazetteer* TGN possui uma coleção com 92% de nomes ambíguos. Dessa forma, diversas estratégias têm sido propostas e descritas nos trabalhos a seguir.

Amitay, Har'El et al [2004] definem dois tipos de ambiguidade para os nomes de lugares, *geo/non-geo* e *geo/geo*. A ambiguidade *geo/non-geo* ocorre quando um nome de lugar possui também um significado não geográfico, como é o caso de “Oliveira”, que é um município em Minas Gerais, mas também pode ser um sobrenome e uma árvore. Já a ambigüidade *geo/geo* ocorre quando dois lugares distintos possuem um mesmo nome, por exemplo, “Rio de Janeiro” pode ser estado e município no Brasil. Outro exemplo é Paris (Texas) e Paris (França).

Volz, Kleb et al. [2007] identificaram três tipos de ambigüidade. A ambigüidade em que nomes iguais podem representar diferentes locais é chamada de *referência múltipla*. Já os casos em que um mesmo lugar é conhecido por diferentes nomes, diz-se que se trata de *nome variante*. O terceiro tipo é a ambigüidade *geoname-non geoname*, semelhante ao identificado por Amitay, Har'El et al [2004], que representa os nomes de lugares que têm outro significado não geográfico.

Diversos métodos para desambiguação de nomes de lugares são propostos para lidar com a ambigüidade e melhorar o desempenho de sistemas de recuperação de informação geográfica [S. E. Overell & Ruger, 2007]. Overell e Ruger [2006] apresentam a aplicação de modelos de co-ocorrência gerados a partir dos artigos da Wikipédia para resolver o problema de desambiguação de nomes de lugares com a utilização de técnicas de aprendizado supervisionado.

Amitay, Har'El et al. [2004] descrevem um componente para desambiguar nomes do sistema Web-a-Where, chamado *geotagger*. Esse componente utiliza um *gazetteer* para encontrar e desambiguar nomes de lugares, como cidades, estados e países. Zong, Wu et al. [2005] propõem regras para desambiguação de nomes (*geo/geo*) de páginas Web através da extração do contexto de informação das páginas e análise das distâncias espaciais entre os lugares candidatos.

Volz, Kleb et al. [2007] descrevem uma abordagem para desambiguação de nomes através de um *ranking* estabelecido entre nomes candidatos reconhecidos em textos com o uso de uma ontologia. Essa ontologia foi criada a partir de dados disponíveis no GeoNames[7] e WordNet[8] e representada com o formato OWL (*Web Ontology Language*), além do esquema em RDF (*Resource Description Framework*). Para avaliação da estratégia proposta, duas coleções de documentos são extraídas da coleção Reuters e classificadas manualmente, sendo a primeira com o objetivo de avaliar a desambiguação de nomes geográficos e a segunda identificar o contexto de páginas.

Pouliquen, Kimler et al. [2006] desambiguam nomes de lugares identificando inicialmente nomes de pessoas e organizações conhecidas, e os nomes restantes são reconhecidos como nomes de lugares através do uso de heurísticas. Gouvêa [2009] propõe uma estratégia para identificar indicadores de localidade no qual são usadas notícias de jornais da Web para a etapa de treino, e com isso não é necessário fazer a seleção manual da coleção. Além disso, uma janela de verificação das relações entre os termos (localidades e indicadores) e fórmulas de pesos são usadas para desambiguar os nomes de lugares.

---

[7] http://www.geonames.org/

[8] http://wordnet.princeton.edu/

Muitas estratégias para desambiguação de nomes geográficos têm usado *gazetteers* como mecanismo de apoio. Entretanto, os tradicionais *gazetteers* somente fornecem listas com nomes de lugares e nomes de lugares alternativos para auxiliar na desambiguação. Essas listas podem ser usadas com o apoio de heurísticas para estabelecer qual dos nomes de lugares ambíguos é o mais provável para um determinado texto. Entretanto, com esses *gazetteers* não é possível realizar inferências com relação aos relacionamentos entre nomes de lugares.

2.3.3 Interpretação de nomes de lugares vagos

Frequentemente, as pessoas usam referências vagas ou aproximadas para nomes de lugares em linguagem natural, como em "sul" ou "norte da Itália". Apesar da provável menção a um lugar definitivo e limitado, o escopo geográfico de tal referência é muito impreciso [Jones & Purves, 2008]. Alguns trabalhos têm demonstrado a necessidade e a interpretação de expressões relacionadas ao espaço geográfico para recuperação de informação geográfica [Delboni et al., 2007]. Geralmente, os *gazetteers* não incluem referências a nomes de lugares vagos e a representação espacial é limitada, o que impede de localizar esses lugares adequadamente.

2.3.4 Indexação espacial e textual

Uma das técnicas para indexar conteúdo de um documento textual é a criação de um arquivo invertido para todas as palavras contidas no documento [Baeza–Yates & Ribeiro-Neto, 1999]. Esse índice permite a associação de cada palavra com a lista dos documentos que a contêm. No caso de referências geográficas, essa idéia pode ser estendida usando-se uma lista de nomes de lugares adicionada a lista de palavras. Um *gazetteer* pode ser facilmente a fonte para essa lista de lugares [Jones & Purves, 2008]. Após a identificação dos nomes de lugares para cada documento, um índice espacial pode ser gerado com o uso de *footprints* (localizações) ou retângulos mínimos envolventes de uma representação geográfica completa.

2.3.5 *Ranking* por relevância geográfica

O *ranking* por relevância geográfica requer uma medida relativa da importância de um documento para uma determinada consulta. Geralmente, os documentos são selecionados de acordo com as ocorrências dos termos da consulta e classificados de acordo com uma medida que considerando as ligações existentes para os documentos candidatos. No caso do *ranking* geográfico, uma associação dos termos da consulta para os lugares encontrados no documento e os *rankings* necessários para combinar critérios geográficos e palavras-chave [Jones & Purves, 2008]. Os tradicionais *gazetteers* não possuem em suas bases outros termos que estão relacionados aos lugares, o que poderia auxiliar na associação entre palavras-chave e nomes de lugares. Entretanto, os *footprints* e representações geográficas, podem ser usados para determinar relacionamentos de proximidade e também influenciar o *ranking* de documentos.

Capítulo 3

Gazetteer Ontológico

O trabalho propõe a criação de uma nova estrutura para os *gazetters* sob a forma de uma ontologia de lugares. Esse capítulo apresenta a proposta de criação da nova estrutura através do diagrama de classe OMT-G para o *gazetteer* Ontológico, também chamado de OntoGazetteer. Descreve o mapeamento do esquema conceitual para um banco de dados geográfico, PostgreSQL. Apresenta como o OntoGazetteer pode ser usado em problemas de recuperação de informação geográfica, além de descrever a interface acesso aos dados do OntoGazetteer.

3.1 Contextualização

Atualmente, os *gazetteers* disponíveis online possuem estruturas muito simples, com apenas três componentes: o nome do lugar, o tipo do lugar e o seu *footprint* (localização geográfica, restrita a um simples par de coordenadas). No Capítulo 2, a estrutura da maioria dos *gazetteers* e suas atuais limitações, que dificultam seu uso em problemas comuns em recuperação de informação geográfica foram descritas.

Souza, Davis-Jr et al. [2005] desenvolveram o Locus, um sistema de localização geográfica que possui um *gazetteer* como principal componente. Os resultados obtidos a partir do desenvolvimento do Locus sugeriram a criação de uma ontologia de lugares, denominada OnLocus [Borges, 2006; Borges et al., 2007]. A ontologia OnLocus descreve as relações espaciais e semânticas entre as localizações, fazendo distinção

entre o lugar real, seu nome e um descritor do lugar. No entanto, a OnLocus foi construída como parte de um esforço para extrair referências indiretas a lugares de páginas da Web, tais como código postal e código de área de telefone [Borges et al., 2007].

O bom desempenho de tais referências indiretas em tarefas de recuperação de informação geográfica mostrou que um *gazetteer* poderia ser muito mais útil se registrasse os vários tipos de relacionamentos que existem entre os lugares e não apenas a topologia de objetos geográficos. Assim poderia incluir outros tipos de relacionamentos semânticos. A implementação de relacionamentos semanticamente mais ricos necessitou da criação de uma estrutura mais flexível para o *gazetteer*, frequentemente encontrada em ferramentas para criação de ontologias, tais como o Protegé[9], que resultou no chamado *gazetteer* ontológico, ou OntoGazetteer.

3.2 Ontologia

Ontologia vem do grego *ontos* (ser) + *logos* (palavra). Na filosofia, este termo é utilizado para descrever tipos de estruturas de entidades, eventos, processos, e relações que existem no mundo real [Breitman, 2006]. Pensando na Web Semântica, Gruber [1992] define ontologia como "uma especificação formal e explícita de uma conceitualização compartilhada". Para a maioria das aplicações na Computação, uma ontologia oferece uma representação do conhecimento de certo domínio, materializada com o objetivo de permitir que máquinas possam trabalhar com o conteúdo semântico de elementos de informação tais como páginas da Web e textos em linguagem natural [Breitman, 2006].

Nos últimos anos foram propostas várias linguagens para construir ontologias. Essas linguagens visam representar conhecimento através da definição de três componentes básicos classes, relacionamentos entre as classes e atributos das classes recomendados pela W3C [Bechhofer et al., 2004]. Atualmente, a linguagem mais usada é a OWL (*Web Ontology Language*) que é uma extensão da RDF (*Resource Description Framework*). Assim, uma ontologia bem construída pode ser um valioso

[9] http://protege.stanford.edu/

mecanismo de apoio a recuperação de informação, pois os dados são semanticamente identificados. Por isso, ontologias têm sido utilizadas na desambiguação de nomes, classificação de documentos, expansão de consultas, entre tantas outras tarefas em RI (Fu, Jones et al. 2005; Fu, Jones et al. 2005; Volz, Kleb et al. 2007; Ping and Yong 2009). Além disso, permitem inferir relações entre os dados e construir um novo conhecimento acerca disso (Abdelmoty, Smart et al. 2007).

O World Wide Web Consortium (W3C) propõe que as ontologias descrevam as classes (ou conceitos) nos vários domínios de interesse, relacionamentos entre as classes e propriedades (ou atributos) que as classes devem possuir [Bechhofer et al., 2004]. Maedche e Staab [2001] definem os elementos de uma ontologia O como: $O: \{C, \mathcal{R}, \mathcal{H}^C, rel, A^\circ\}$, onde C é o conjunto de conceitos ou classes, $\mathcal{R}$ é o conjunto de relacionamentos, $\mathcal{H}^C$ um relacionamento direto, rel é uma função que relaciona conceitos de modo não-taxonômico e A° é um conjunto de axiomas expresso em uma linguagem lógica apropriada.

As linguagens usadas para definir ontologias possuem limitações na representação de determinados domínios. Por exemplo, a criação de ontologias de contexto geográfico não é uma tarefa simples, visto que a OWL não permite representar informações espaciais, bem como não possui regras específicas para deduzir relacionamentos e restrições de integridade espaciais (Abdelmoty, Smart et al. 2007). Diante dessas limitações, Jones, Alani et al. [2001] implementam uma ontologia apoiada em *gazetteers* e tesauros. Abdelmoty, Smart et al. [2007] propõem o uso de dois frameworks, além da linguagem OWL, para construir e manter ontologias de lugares. Já Huang e Deng [Huang & Deng, 2009] propõem o uso de SWRL para expressar regras espaços-temporais e construir geo-ontologias. Lopez-Pellicer, Silva et al. [2010] criam uma ontologia geográfica de Portugal a partir do desenvolvimento de um vocabulário, *Geo-Net*, implementado em RDF e composto por um conjunto de classes e propriedade que usam GML (*Geography Markup Language*), WKT (*Well-Known Text*), entre outros padrões propostos.

Esse trabalho apresenta uma proposta de especificação e implementação de um *gazetteer*, ou dicionário de nomes de lugares, concebido com base na estrutura usual de uma ontologia. A idéia é estabelecer uma correspondência entre elementos de ontologias e o conteúdo do *gazetteer*, e a partir daí propor inovações úteis para a

função que os *gazetteers* exercem em sistemas de recuperação de informação. A partir desta correspondência, um conjunto de requisitos para o conteúdo do *gazetteer* ontológico, que vai muito além do que é geralmente incluído. Nossa intenção é permitir que pesquisadores usem o dicionário geográfico como um instrumento para situações em que há a necessidade de reconhecer o contexto geoespacial de documentos. O *gazetteer* ontológico ou OntoGazetteer oferece recursos que permitem que ele seja usado em problemas de recuperação de informação geográfica, como a detecção de referências geográficas, desambiguação nome do lugar, a interpretação de nomes de lugares vagos, a indexação espacial e textual, e *ranking* de relevância geográfica.

Tabela 3-1 – Correspondência entre ontologia e elementos do *gazetteer*

Relacionamento ontológico	Definição	Equivalente no *gazetteer*	Exemplo
Sinonímia	Um termo X tem quase o mesmo significado que o termo Y.	Nomes alternativos, apelidos, nomes históricos, abreviações, variações de grafia	*Belo Horizonte:* BH, Belô, B. Horizonte, Cidade de Minas
Homonímia	O termo X tem a mesma grafia que o termo Y, porém tem significado diferente.	Nomes ambíguos: lugares (do mesmo tipo ou de tipos diferentes) com nomes coincidentes, lugares com nomes de outras coisas	São Paulo (município) e São Paulo (time de futebol); municípios de Viçosa (MG) e Viçosa (AL); Vitória (ES) e o substantivo *vitória*
Hiperonímia	Um termo X tem significado mais amplo do que o termo Y.	Lugar em nível superior em uma hierarquia espacial	Minas Gerais em relação a Belo Horizonte.
Hiponímia	Um termo X tem significado mais restrito do que o termo Y.	Lugar em nível inferior em uma hierarquia espacial	Bacia do Rio Tietê em relação à bacia do Rio Paraná
Associação	O termo X está associado a um termo Y, isto é, existe um relacionamento semântico entre eles.	Relacionamento semântico entre lugares	Municípios ao longo da BR-040; Municípios produtores de soja; rios da Amazônia; cidades históricas; estâncias hidrominerais

Inicialmente, estabelecemos na Tabela 3-1 uma correspondência direta entre as estruturas de uma ontologia e elementos necessários para um *gazetteer*. Inicialmente, é feita a correspondência entre Conceito (ontologia) e Lugar (*gazetteer*). Da mesma

forma que um conceito pode ter várias designações em linguagem natural, um lugar pode ter vários nomes, com variações culturais, temporais e linguísticas. Os relacionamentos entre conceitos em uma ontologia têm também uma correspondência com relacionamentos espaciais e semânticos entre os lugares. Termos sinônimos em uma ontologia correspondem a nomes alternativos para o mesmo lugar no *gazetteer*. Hiponímia e hiperonímia correspondem aos relacionamentos hierárquicos de subdivisão espacial, tais como os que existem entre um estado e seus municípios. O mesmo pode ser dito da meronímia (relacionamento "todo-partes", no caso subdivisões políticas ou físicas, como sub-bacias de uma bacia hidrográfica). Outros tipos de associações entre termos em uma ontologia podem ser mapeados para relacionamentos espaciais (por exemplo, de vizinhança ou proximidade) ou semânticos (por exemplo, o relacionamento entre cidades que são capitais de estados). Foi incluído também relacionamentos entre lugares e termos, tanto para denotar ambiguidade (lugares que têm nomes de pessoas ou coisas) quanto para registrar associações (termos que estão associados a lugares).

Para permitir a representação dos lugares, seus nomes, sua localização geográfica, relacionamentos dos diversos tipos entre lugares, bem como termos associados e outros detalhes, implementamos o OntoGazetteer com base em um banco de dados, cujo projeto conceitual é apresentado a seguir.

3.3 Esquema Conceitual

O esquema proposto para o OntoGazetteer foi modelado através de um diagrama OMT-G [Machado, Alencar, Campos-Jr, & Davis-Jr, 2010]. OMT-G é um modelo de dados para representar aplicações geográficas proposto por [Borges, Davis-Jr, & Laender, 2001]. Esse modelo foi usado por permitir a modelagem das diversas representações geográficas para os lugares, tais como `Area` ou `Polígono`, `Ponto`, `Linha`, entre outras.

A Figura 3-1 apresenta o esquema proposto. O esquema representa todos os nomes de lugares através da classe `Lugar`. A classe `Lugar_Alternativo` mantém nomes alternativos, abreviaturas, siglas, nomes anteriores, apelidos, entre outras

variações para os nomes de lugares. Cada lugar pertence a um tipo de lugar representado pela classe `Tipo_Lugar`. Inicialmente os tipos de lugares foram definidos com base no *thesaurus* da ADL.

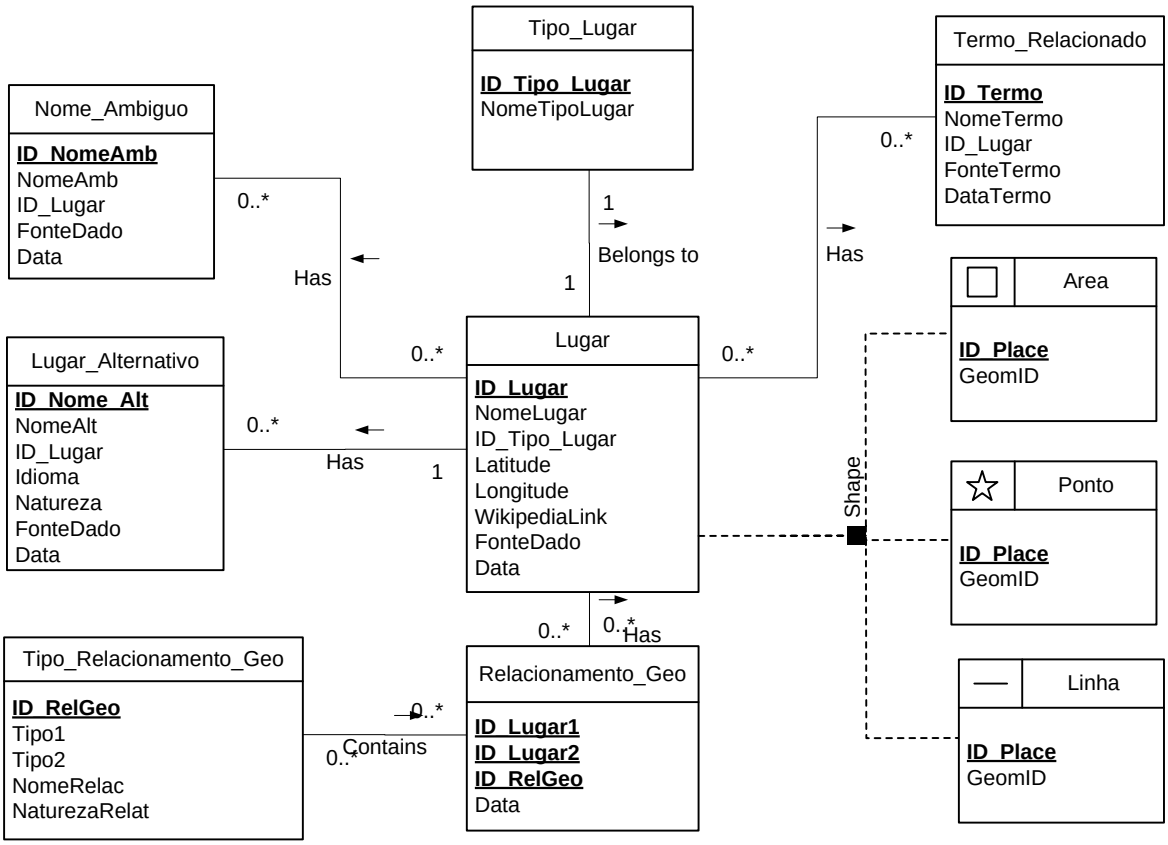


Figura 3-1 - Esquema Conceitual para o Gazetteer Ontológico

Os diversos relacionamentos entre os lugares são mantidos na classe `Relacionamento_Geo`. Sendo que dois lugares podem ter vários relacionamentos entre eles, diferenciado pelo tipo, que é representado pela classe `Tipo_Relacionamento_Geo`. Essa classe permite que o *gazetteer* grave e use uma grande quantidade de conexões semânticas e geográficas entre os lugares. Atualmente os tipos de relacionamentos usados são apenas o hierárquico e de vizinhança. No entanto, essa estrutura permite a criação de relacionamentos semânticos, como por exemplo, município que uma rodovia ou rio atravessam, ou cidades que são capitais brasileiras, ou municípios em Minas Gerais que possuem

grande atividade agropecuária. Além desses relacionamentos é possível criar outros diversos relacionamentos.

A desambiguação de nomes é apoiada pela classe `Nome_Ambiguo`, que mantém nomes ambíguos já conhecidos. A classe `Termo_Relacionado` armazena uma lista de termos relacionados aos lugares, que além de apoiar a desambiguação podem apoiar outras aplicações para recuperação de informação geográfica. Uma das fontes utilizada para identificar esses termos foi a Wikipédia [Alencar et al., 2010], por isso, a classe `Lugar` mantém um atributo para a URL de sua entrada na Wikipédia. Por exemplo, a classe `Termo_Relacionado` pode conter o termo "acarajé" associado ao lugar "Bahia".

Finalmente, cada lugar pode ter uma ou mais representações geográficas, representadas pelas classes `Ponto`, `Linha` ou `Polígono` [Davis-Jr & Laender, 1999]. Dessa forma um lugar pode ter uma representação geográfica completa, Lugares representados por geometrias complexas também serão representados através de um ponto, como nos atuais *gazetteers*.

3.4 Construção do Banco de Dados

As classes do *gazetteer* ontológico foram mapeadas para tabelas em um SGBD (Sistema Gerenciador de Banco de Dados) que permite o armazenamento de dados espaciais, portanto um SGBD geográfico. A escolha para o OntoGazetteer foi o PostgreSQL 8.4[10] por ser um banco de dados distribuído sob uma licença de software livre (BSD) e possuir um módulo específico para tratar dados espaciais, o PostGIS[11].

A construção do banco de dados foi feita, em geral de forma manual e buscou abranger dados do Brasil como um todo até o nível de bairros. Diversas fontes de dados geográficos foram usadas, tais como o Instituto Brasileiro de Geografia e Estatística (IBGE), o *gazetteer* da *Alexandria Digital Library* (ADL), dados do projeto GeoMinas, da Prefeitura de Belo Horizonte, da Wikipédia[12], do *gazetteer* do Locus [Souza et al.,

10 http://www.postgresql.org.br/

11 http://postgis.refractions.net/

12 http://pt.wikipedia.org/

2005], do Centro de Análises de Sistemas Navais[13] (CASNAV), entre outras prefeituras. A Tabela 3-2 apresenta as quantidades de registros carregados até o momento para cada uma das tabelas do banco de dados e suas respectivas fontes de dados.

Tabela 3-2 - Quantidade de registros do OntoGazetteer

Tabela	# registros	Fontes de dados
Lugar	151.029	IBGE, CASNAV, GEOBASES, Prefeitura Rio, Prodabel
Tipo_Lugar	21	ADL, Prodabel
Lugar_Alternativo	74.191	Wikipédia, IBGE, próprio banco de dados
Termo_Relacionado	247.590	Wikipédia
Relacionamento_Geo	198.957	Próprio banco de dados
Tipo_Relacionamento_Geo	39	Próprio banco de dados

A atualização da tabela de `Relacionamentos Geográficos` é feita a partir dos próprios dados geográficos sobre lugares, à medida que novos dados são carregados no banco de dados. Dessa forma, *scripts* em PHP e Java foram desenvolvidos para tentar diminuir o trabalho para manter o OntoGazetteer. O Apêndice A apresenta dois scripts utilizados para inserir os relacionamentos geográficos hierárquicos e por vizinhança.

3.5 Aplicações do OntoGazetteer

Os *gazetteers* podem ser usados como mecanismos para solucionar muitos dos problemas de recuperação de informação geográfica como descrito no Capítulo 2. Entretanto, como o OntoGazetteer possui uma estrutura que visa diminuir algumas das limitações dos *gazetteers* tradicionais, pode fornecer um suporte melhor para resolver esses e outros problemas em GIR. A seguir são descritas algumas das possíveis contribuições do OntoGazetteer para resolver problemas de GIR.

13 https://www.casnav.mar.mil.br/principal.php

O OntoGazetteer mantém listas de nomes de lugares oficiais e alternativos, que facilitam a identificação de nomes de lugares candidatos encontrados em textos. Além disso, mantém os relacionamentos entre os lugares, que são identificados a partir de um esforço inicial. Essa lista de relacionamentos entre lugares pode ser usada tanto para identificar as reais referências a lugares geográficos entre os termos candidatos, como auxiliar na desambiguação desses nomes de lugares. O OntoGazetteer também mantém informações sobre as hierarquias espaciais e lugares adjacentes, e com isso, é possível inferir a co-ocorrência de lugares relacionados e desconsiderar outros nomes lugares candidatos. O contexto de um documento pode se referir a algum nível superior da hierarquia espacial, por exemplo, um texto que menciona várias cidades em um mesmo estado, tem como contexto real o próprio estado.

A desambiguação de nomes com o OntoGazetteer também pode ser feita com o uso das listas de nomes de lugares ambíguos, nomes alternativos e termos relacionados. Essas listas podem ser usadas com o apoio de heurísticas para estabelecer qual dos nomes de lugares ambíguos é o mais provável para um determinado texto. Os relacionamentos entre os nomes de lugares podem ajudar na desambiguação de nomes da seguinte forma. Um relacionamento fraco entre um nome de lugar com os outros elementos encontrados no mesmo texto (outros nomes de lugares, outros termos de linguagem natural) podem indicar um termo ambíguo e ser desconsiderado.

A interpretação de nomes de lugares vagos e a indexação espacial e textual pode ser feita com uso da representação geográfica completa disponível no OntoGazetteer. A interpretação de nomes de lugares vagos pode usar inferências para identificar uma subdivisão do lugar mencionado usando pistas associadas a expressões de linguagem natural. Já a indexação espacial e textual que usar uma representação geográfica completa poderá futuramente recuperar os documentos usando relacionamentos espaciais, tais como proximidade e de contido.

Como o OntoGazetteer mantém listas de termos relevantes, uma estratégia para *ranking* pode ser determinar o quanto específico termos de consultas estão relacionados a lugares, ajudando assim a diminuir os resultados e atribuir mais importância aos documentos em que ambos os termos e os lugares estão relacionados com a consulta. A partir de *footprints* e representações geográficas, relacionamentos de proximidade podem ser determinados, podendo também influenciar o *ranking* de documentos.

3.6 Interface do OntoGazetteer

Uma aplicação Web foi desenvolvida em Java para disponibilizar acesso seguro pela Web aos dados do *gazetteer* ontológico. Dessa forma, o usuário pode realizar uma pesquisa simples por um nome de lugar com ou sem filtro do tipo de lugar. A Figura 3-2 apresenta a tela inicial da interface disponível para o usuário através da URL http://www.greenwich.lbd.dcc.ufmg.br:8080/ontogazetteer/.

Figura 3-2 - Interface Inicial do OntoGazetteer

A Figura 3-3 apresenta o resultado para uma busca pelo município de "Ouro Preto". O resultado é dado sob a forma de um relatório, no qual os diversos lugares relacionados ao nome buscado são apresentados em forma de *links*. Ao clicar nos *links* é uma nova busca é realizada para o lugar selecionado e um novo relatório será apresentado.

1: OURO PRETO (MUNICIPIO)
1:1 VIZINHO DO MUNICIPIO OURO BRANCO
1:2 VIZINHO DO MUNICIPIO PIRANGA
1:3 VIZINHO DO MUNICIPIO SANTA BARBARA
1:4 VIZINHO DO MUNICIPIO CATAS ALTAS DA NORUEGA
1:5 VIZINHO DO MUNICIPIO MARIANA
1:6 VIZINHO DO MUNICIPIO CONGONHAS
1:7 VIZINHO DO MUNICIPIO BELO VALE
1:8 CONTIDO NA REGIAO SUDESTE
1:9 CONTIDO NA MICRORREGIAO OURO PRETO
1:10 VIZINHO DO MUNICIPIO ITABIRITO
1:11 CONTIDO NO ESTADO DE MINAS GERAIS
1:12 CONTIDO NA MESORREGIAO METROPOLITANA DE BELO HORIZONTE
1:13 VIZINHO DO MUNICIPIO ITAVERAVA
1:14 VIZINHO DO MUNICIPIO MOEDA
1:15 NOME ANTERIOR : VILA RICA DE ALBUQUERQUE
1:16 NOME ANTERIOR : VILA RICA

Figura 3-3 - Relatório para pesquisa por 'Ouro Preto'

Como resultado para a mesma busca pelo município de "Ouro Preto", um mapa com a localização geográfica também é apresentado, como mostra a Figura 3-4. A API do Google foi usada para criação dos mapas na interface Web. Dessa forma, é possível aproximar ou distanciar do ponto localizado no mapa, o que facilita a identificação dos lugares relacionados (contidos, vizinhos, alternativos, entre outros) apresentados no relatório com o resultado para a pesquisa.

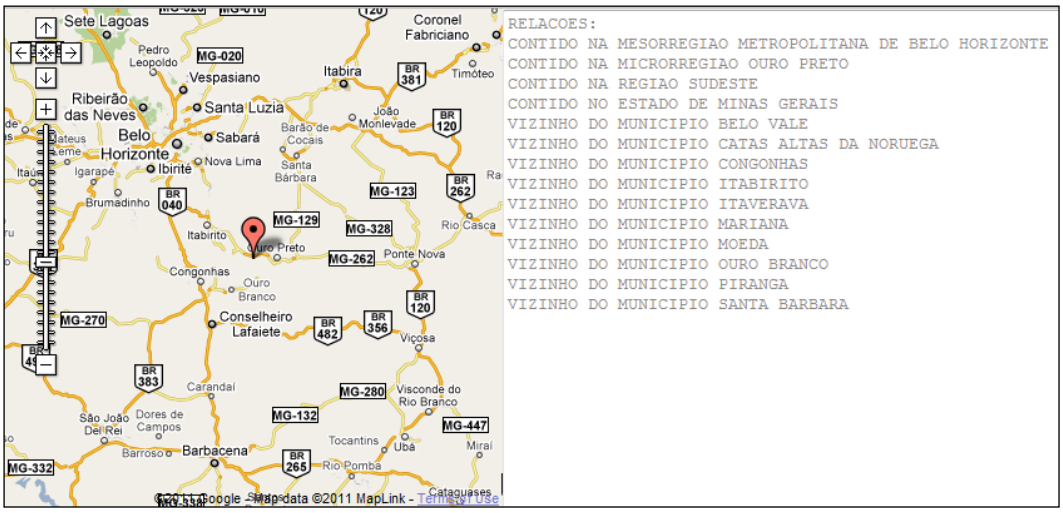


Figura 3-4 - Mapa para pesquisa por 'Ouro Preto'

Além da pesquisa simples o usuário pode obter os dados do OntoGazetteer na forma de um arquivo XML (*Extensible Markup Language*) com os resultados para uma determinada consulta realizada por um usuário. Entretanto, pretende-se disponibilizar os dados do OntoGazetteer sob a forma de serviços Web, para permitir um uso melhor e mais ativo desses dados. Um serviço Web (ou *Web service*) é um recurso que possibilita executar funções e realizar consultas através de um protocolo de comunicação baseado na Web, de forma independente da tecnologia usada para gerenciar e armazenar os dados. A implementação dos serviços Web faz com que as requisições e as respostas dos serviços sejam codificados em XML, permitindo que tanto um usuário, quanto um aplicativo podem acessar dados de seu interesse no OntoGazetteer de forma segura e padronizada através da Web, conforme mostra a Figura 3-5.

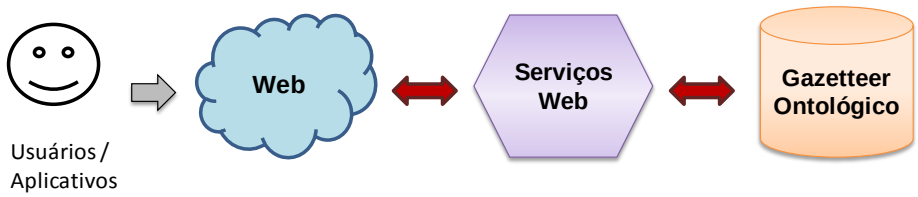


Figura 3-5 - Funcionamento dos serviços Web

Capítulo 4

Avaliação do OntoGazetteer

Esse capítulo apresenta a aplicação do *gazetteer* proposto através de dois estudos de caso. Foram escolhidas duas aplicações de Recuperação de Informação Geográfica para avaliar o uso do OntoGazetteer: (1) detecção e inferência do contexto geográfico em textos e (2) desambiguação de nome de lugares. Em cada caso, foram preparados e executados experimentos em que a informação disponível no *gazetteer* ontológico fornece meios para a resolução do problema.

4.1 Contextualização

Os estudos de caso usaram, como fontes de dados, textos de notícias encontrados na Web. Normalmente, os textos jornalísticos contêm um ou mais locais relacionados com os fatos descritos, e isso faz parte da técnica de escrita da notícia. Para os dois estudos, o primeiro passo realizado foi a coleta de notícias e a extração do título e o corpo da notícia. Com isso, foram extraídos das páginas Web apenas os dados que realmente eram importantes para a realização dos experimentos.

O primeiro estudo de caso apresenta uma estratégia para detecção e inferência do contexto geográfico em notícias apoiado com o OntoGazetteer. Já o segundo estudo de caso apresenta uma estratégia para desambiguar nomes de lugares também apoiada pelo OntoGazetteer. Os experimentos realizados foram avaliados por voluntários e mostraram bons resultados, o que valida o uso do OntoGazetteer. Essas estratégias e

resultados são detalhados a seguir.

4.2 Estudo de Caso: Detecção e Inferência do Contexto Geográfico em Textos

Esse estudo de caso tem como objetivo detectar a ocorrência de nomes de lugares e inferir o contexto geográfico geral do texto, que pode estar explicitamente indicado no mesmo ou ter que ser inferido a partir das evidências encontradas nele. Por isso, os nomes de lugares detectados no texto da notícia foram denominados *lugares explícitos*. Já os lugares inferidos a partir desses lugares explícitos e seus relacionamentos foram denominados *lugares implícitos* [Machado et al., 2010].

Jovem é encontrado morto em lote vago no Bairro Ouro Preto

Estado de Minas -
Publicação: 03/08/2010 08:45 Atualização: 03/08/2010 08:58

A polícia investiga um homicídio ocorrido no fim da noite de segunda-feira no Bairro Ouro Preto, Região da Pampulha. Jonny Moreira da Silva, de 23 anos, foi encontrado morto com dois tiros na cabeça em um lote vago na Rua Luiz Lopes. Segundo militares do 34º Batalhão, o rapaz estava com cinco pedras de crack.

A autoria e a motivação do crime são desconhecidas, mas a polícia suspeita de envolvimento com tráfico de drogas. O corpo do jovem foi levado para o Instituto Médico Legal (IML). O registro foi feito na Seccional Noroeste.

Figura 4-1 - Exemplo de Notícia Estudo de Caso 1 (fonte: Uai, 3/8/2010)

A Figura 4-1 apresenta um exemplo de uma notícia do jornal Web *Uai Minas*[14], publicada em 3 de agosto de 2010. Nessa notícia são identificados dois nomes de lugares explícitos, "Ouro Preto" e "Pampulha". De acordo com o conteúdo do *gazetteer* ontológico, "Ouro Preto" é um nome ambíguo, pois há cinco lugares com o mesmo nome. No entanto, esse texto se refere a um bairro de Belo Horizonte e não à famosa cidade histórica. Através do processo proposto tanto a cidade quanto o bairro são

[14] http://www.uai.com.br/htmls/app/noticia173/2010/08/03/noticia_minas,i=172053/index.shtml

obtidos como possíveis candidatos, mas a partir dos relacionamentos entre esses nomes de lugares, outros nomes de lugares são inferidos como "Belo Horizonte" (cidade); "Metropolitana de Belo Horizonte" (microrregião); "Ouro Preto" (microrregião); "Metalúrgica e Campos das Vertentes" (macrorregião) e "Minas Gerais" (Estado).

Esse texto também inclui uma referência à rua “Rua Luiz Lopes”, mas atualmente o OntoGazetteer não contém esse tipo de lugar. Além disso, indicadores de localidade como “Bairro” e “Região” que precedem os nomes de lugares não foram usados no processo proposto. Entretanto, para o próximo estudo de caso esses indicadores foram usados para desambiguar nomes de lugares.

4.2.1 Criação da coleção de notícias

A coleta de notícias para a criação dessa coleção foi feita de junho a agosto de 2010 (Tabela 4-1). Um coletor específico para cada uma das fontes de notícias online foi desenvolvido com o uso de XPath para extrair e armazenar o título e corpo da notícia. Além disso, foram coletadas apenas notícias do estado de Minas Gerais. Para assegurar isso, as notícias foram obtidas a partir de seções locais ou seções estaduais referentes a Minas Gerais das fontes de notícias.

Tabela 4-1 - Fontes de Notícias Web

Jornal Web	Coleta	Web site	Nome da seção do jornal	# Docs
Globominas	06 a 08/2010	http://globominas.globo.com/	Notícias gerais	139
O Tempo	06 a 08/2010	http://www.otempo.com.br/	Últimas notícias - Cidades	75
Uai	06 a 08/2010	http://www.uai.com.br/	Minas	71
Terra	06 a 08/2010	http://www.terra.com.br	Últimas notícias - Brasil	11
			Total	**296**

4.2.2 Detecção e inferência do contexto geográfico em textos

Após a coleta de notícias, um pré-processamento foi realizado para retirar *stopwords*

(exceto "de ", "da (s)", "do (s)", que são bem comuns em nomes de lugares do Brasil). Conforme descrito no Capítulo 2, *stopwords* são palavras que podem ser consideradas irrelevantes para o contexto do sistema, nesse caso, o contexto geográfico. A lista de *stopwords* removidas está disponível no Apêndice B-1.

```
"(" + "MG-[0-9]+)|"+ "(BR-[0-9]+)|"+"(LMG-[0-9]+)|" +"(MGT-[0-9]+)|" +
"(([A-ZÁÀÃÂÄÉÊÍÓÔÕÖÚÜÑ])([A-ZÁÀÃÂÄÉÊÍÓÔÕÖÚÜÑ]+))|" +
"(([A-ZÁÀÃÂÄÉÊÍÓÔÕÖÚÜÑ])([a-záàãâäéêíóôõöúüçñ]+)( ([A-
ZÁÀÃÂÄÉÊÍÓÔÕÖÚÜÑ])([a-záàãâäéêíóôõöúüçñ]+))+)|"+
"(([A-ZÁÀÃÂÄÉÊÍÓÔÕÖÚÜÑ])([A-ZÁÀÃÂÄÉÊÍÓÔÕÖÚÜÑ]+|[a-
záàãâäéêíóôõöúüçñ]+))"+ ")";
```

Figura 4-2 – Expressões regulares

Ainda na etapa de pré-processamento, termos candidatos a nomes de lugares foram identificados e extraídos do texto da notícia, usando expressões regulares para identificar nomes próprios, compostos ou não. A Figura 4-2 apresenta as expressões regulares usadas para identificar esses termos candidatos a nomes de lugares, portanto, nomes próprios e nomes de rodovias de Minas Gerais.

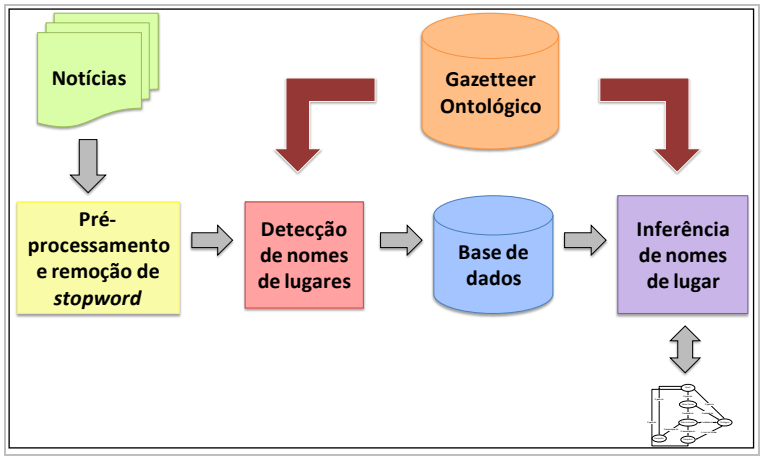


Figura 4-3 – Processo para detecção e inferência de nomes de lugares

O reconhecimento e identificação do tipo dos nomes de lugares encontrados na notícia foi apoiada pelo OntoGazetteer, como mostra a Figura 4-3. Dessa forma, um

casamento simples foi realizado entre os termos candidatos e os nomes de lugares e os nomes de lugares alternativos presentes no *gazetteer*. Os termos candidatos, agora já identificados como nomes de lugares, foram armazenados em um banco de dados. Posteriormente, essas instâncias de nomes de lugares, também denominados lugares explicítos, juntamente com as instâncias da classe `Relacionamento_Geo`, foram utilizadas para inferir referências implícitas a outros lugares através dos relacionamentos entre os nomes de lugares.

O Algoritmo 4-1 apresenta a função *RecuperarContextos*, que recupera o contexto geográfico para cada uma das notícias da coleção que ainda não tenha sido processada. O algoritmo começa com a recuperação de uma lista com todas as notícias, que ainda não tenham sido processadas, através da função *RecuperarListaNoticias*. Em seguida, os termos candidatos a nomes de lugares de cada uma das notícias da *listaDeNoticias* é recuperado atraves da função *RecuperarListaTermos.* Esses termos candidatos são então comparados com os nomes de lugares da classe `Lugar` e com os nomes alternativos para lugares da classe `Lugar_Alternativo` do *gazetteer* Ontológico. Após esse passo, a notícia é então atualizada para o status de processada e uma função para recuperar os nomes de lugares implícitos é chamada.

```
1   Função RecuperarContextos()
2      {
3          listaDeNoticias = Função RecuperarListaNoticias;
4          Para cada (Noticia na listaDeNoticias)
5          {
6              listaDeTermos = Função RecuperarListaTermos;
7              Para cada (Termo na listaDeTermos)
8              {
9                  Se (noticia != 'PROCESSADA')
10                     {
11                         listaDeLugares = Função RecuperarLugar;
12                         listaDeLugares = Função RecuperarLugarAlternativo;
13                     }
14                 }
15                 Função AtualizarStatusNoticia('PROCESSADA');
16                 Função RecuperaLugarImplicito(Noticia);
17         }//fim Para cada notícia
18     } //fim do Função Recuperar Contexto
```

Algoritmo 4-1 – Função Recuperar Contextos

O Algoritmo 4-2 apresenta a função *RecuperaLugarImplícito*, que é responsável por identificar os nomes de lugares implícitos a partir dos nomes de lugares já identificados (explícitos). Foram identificados e usados apenas os nomes de lugares mais elevados na hierarquia territorial.

```
1  Função RecuperaLugarImplicito(Notícia)
2          {
3              listaDeLugaresImplicitos;
4              listaDeLugaresExplicitos;
5              listaDeLugaresExplicitos = Função RecuperaLugaresExplicitosBD(Noticia);
6
7              Para cada lugar explícito()
8              {
9                  lstaLugaresImplicitos = RecuperarLugarImplicitoBD();
10             }
11             GravarListaDeLugares();
12         }
```

Algoritmo 4-2 – Função Recupera Lugar Implícito

A Figura 4-4 apresenta o grafo de referências explícitas e implícitas resultante para a notícia da Figura 4-1. Dos seis nomes de lugares identificados inicialmente como candidatos, o algoritmo foi capaz de selecionar os corretos e, em seguida inferir dois nomes de lugares a partir do relacionamento, Belo Horizonte e Minas Gerais.

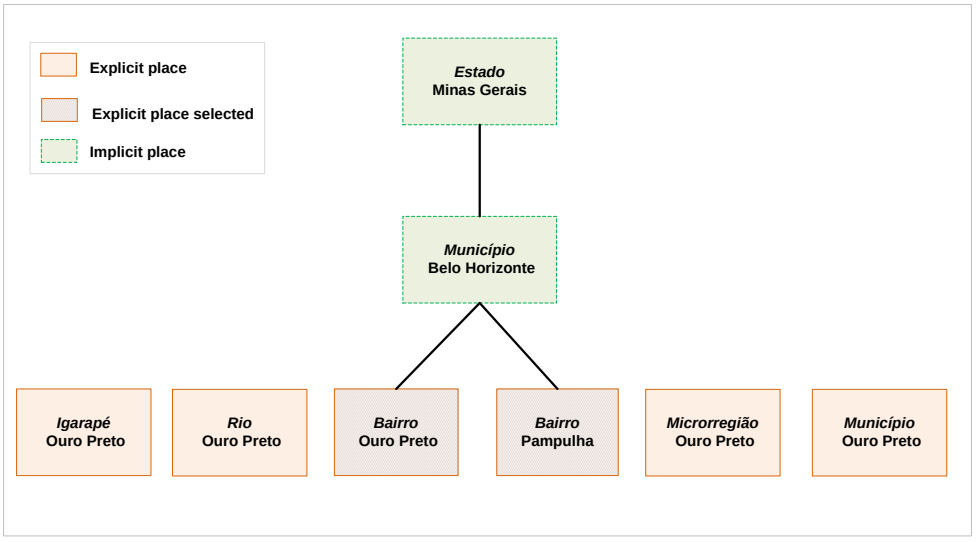


Figura 4-4 - Grafo de inferência de referências explícitas e implícitas

4.2.3 Avaliação e resultados dos experimentos

A validação do processo proposto para detecção e inferência do contexto geográfico em notícias foi feita através de uma verificação manual de cada uma das notícias por um grupo de voluntários, composto por pessoas com diferentes níveis de conhecimento, mas em geral foram escolhidos voluntários formados em Ciência da Computação ou áreas afins. Uma interface Web foi desenvolvida para facilitar a verificação feita pelo voluntário. Dessa forma, o voluntário poderia avaliar quantas notícias desejasse em qualquer local e momento. Além disso, um tutorial foi disponibilizado para explicar o objetivo da pesquisa e o funcionamento da ferramenta Web para avaliação.

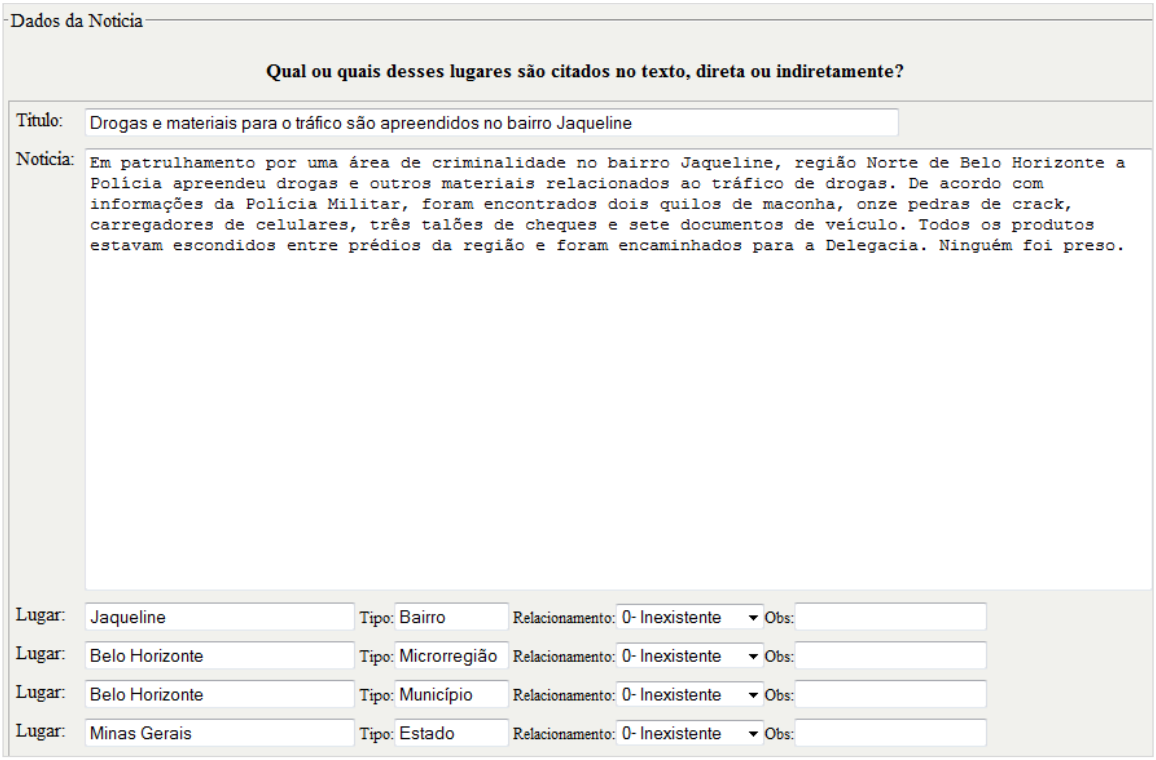


Figura 4-5 – Interface Web para avaliação dos voluntários

A Figura 4-5 apresenta a interface para avaliação de uma determinada notícia. Essa interface apresenta, para cada notícia selecionada, o título, o corpo da notícia e a lista de lugares explícitos e implícitos gerados a partir do processo proposto. Nessa interface (Figura 4-5), os voluntários deveriam responder para cada lugar da lista de

lugares o grau do relacionamento com a notícia utilizando uma escala de cinco níveis: (0) não relacionado, (1) relacionamento muito fraco, (2) relacionamento fraco, (3) indiferente, (4) relacionamento forte e (5) relacionamento muito forte. Caso o voluntário não tivesse conhecimento pessoal sobre determinado lugar, poderia abandonar a questão e analisar outra notícia. Além disso, para cada lugar avaliado, o voluntário poderia colocar alguma observação ou indicar dificuldades e casos especiais que pudessem ser usados para analisar os resultados dos experimentos e promover melhoramentos no processo.

Para facilitar a avaliação dos resultados dos experimentos, a escala com os níveis de relacionamentos foi agrupada em apenas três níveis: (0) não relacionado, (1) relacionamento fraco, (2) relacionamento forte, (3) indiferente. Das 296 notícias coletadas (Tabela 4-1), foram extraídos nomes de lugares de apenas 267 notícias e, portanto, apenas essas foram utilizadas no experimento como um todo. Além disso, em média, em cada notícia foram identificados 8,46 nomes de lugares, explícitos ou implícitos.

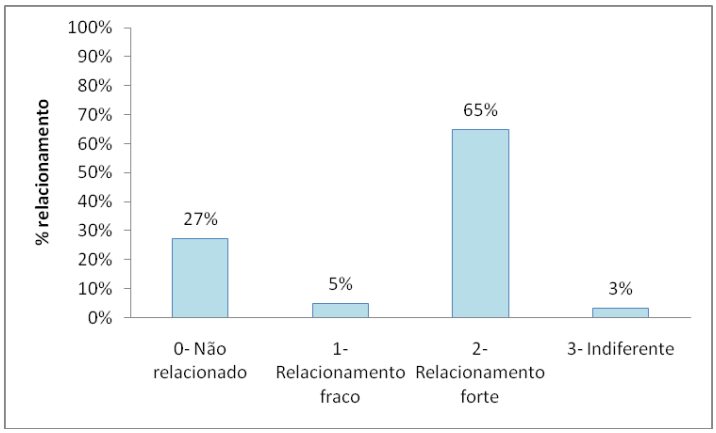


Figura 4-6 – Gráfico com grau de relacionamento

Tabela 4-2 - Avaliação do grau do relacionamento

Grau do relacionamento	#	%
0- Não relacionado	609	27,1
1- Relac. Fraco	106	4,7
2- Relac.forte	1.455	64,8
3- Indiferente	74	3,4
Total	**2.244**	**100,0**

Os voluntários verificaram 100% das 267 notícias que possuíam nomes de lugares. Foram avaliados 2.244 relacionamentos entre lugares e notícias, dos quais 72% foram considerados válidos (Tabela 4-2), a maioria dos quais considerados relacionamentos fortes pelos voluntários (Figura 4-6).

A Tabela 4-3 apresenta uma avaliação de acordo com o tipo de lugar. As colunas com a média e desvio padrão referem-se a um grau de parentesco, sendo assim, o nível (3) indiferente foi desconsiderado para esses cálculos. As médias que se aproximam de 2,0 indicam que existe uma forte relação entre os nomes de lugares de um determinado tipo e suas notícias. Observou-que os nomes de lugares cujos nomes eram mais conhecidos pelos voluntários tiveram uma melhor classificação na avaliação. Além disso, os voluntários tiveram mais dificuldade para avaliar os tipos de lugar que pertencem a uma hierarquia territorial administrativa, tais como macrorregião, mesorregião e microrregião. Esses tipos de lugares foram incluídos em geral como lugares implícitos, já que esses nomes quase nunca são mencionados em textos jornalísticos. Um exemplo muito recorrente, e que poucos voluntários, aparentemente, sabiam que o município de Belo Horizonte, em Minas Gerais, pertence à macrorregião chamada de "Metalúrgica e Campos das Vertentes".

Tabela 4-3 – Avaliação de acordo com o tipo de lugar

Tipo de lugar	Total	Refs. Implícitas	Média	Desvio padrão
Macrorregião	184	171	0,36	0,75
Microrregião	298	129	0,98	0,94
Rio	2	0	1,00	1,41
Mesorregião	246	0	1,17	0,95
Município	446	35	1,50	0,85
Estado	341	233	1,61	0,75
Bairro	546	0	1,73	0,67
Rodovia	107	0	1,91	0,42

Ainda na Tabela 4-3 é possível verificar que uma grande quantidade de bairros foi identificada. Isso ocorre porque é elevado o número de menções explícitas a bairros em notícias locais (ou urbanas), reforçando a idéia de que *gazetteers* devem abranger detalhes de lugares intra-urbanos. Rodovias também foram citadas em algumas notícias, com isso supõe-se que as relações entre os lugares ao longo de uma rodovia

podem ser semanticamente muito importantes. Além disso, foi identificada uma alta incidência de referências implícitas a Estados, algo que era já esperado, já que as fontes de notícias foram seções estaduais e locais de jornais online. Como ponto negativo, o pequeno número de referências a rios indica que a cobertura do OntoGazetteer de recursos hídricos é atualmente deficiente. Por outro lado, existe muita repetição de nomes de rios na hidrografia nacional, é de se esperar que o reconhecimento de nomes de corpos d'água seja muito complicado devido à ambigüidade.

4.3 Estudo de Caso: Desambiguação de Nomes de Lugares

Esse estudo de caso tem como objetivo desambiguar nomes de lugares em notícias da Web. Para isso, algumas das estratégias usadas no estudo anterior, como detecção de nomes de lugares e inferência de contexto geográfico de notícias da Web através dos lugares implícitos identificados a partir dos relacionamentos dos nomes de lugares extraídos, continuam sendo usadas.

Carga espalhada na Via Expressa deixa trânsito lento em Contagem

Estado de Minas -
Publicação: 22/07/2010 14:14

Um acidente inusitado deixou o trânsito lento no Bairro Eldorado, em Contagem, Região Metropolitana de BH. Chapas de aço caíram de um caminhão que passava por cima de um viaduto na Avenida João César de Oliveira.

O material caiu sobre a Via Expressa interditando o trânsito no sentido Centro de Contagem. Segundo a Transcon, o trânsito ficou lento na região e apenas a faixa da direita está liberada para a passagem de veículos. As chapas já estão sendo recolhidas. Niguém ficou ferido.

Figura 4-7 - Exemplo de Notícia Estudo de Caso 2 (fonte: Uai - Minas, 22/07/2010)

A Figura 4-7 apresenta um exemplo de uma notícia do jornal Web *Uai Minas*[15], publicada em 22 de julho de 2010. Nessa notícia foram encontrados nomes de lugares no título e corpo da notícia, marcados na figura (Contagem, Eldorado, Metropolitana de BH). A presença desses nomes e as relações espaciais entre eles indicam que a notícia implicitamente trata de outros lugares também, por exemplo, o estado de Minas Gerais e a microrregião Belo Horizonte. Estes lugares não estão explícitos no texto, mas podem ser identificados a partir das relações que têm com os lugares explicitamente mencionados.

No entanto, os lugares explicitamente mencionados no texto podem ser ambíguos. Por exemplo, o nome Eldorado na Figura 4-7 está associado a um bairro na cidade de Contagem, Minas Gerais, mas também a outros lugares nos estados do Espírito Santo e Mato Grosso do Sul. Ao considerar outros nomes que ocorrem no texto e os relacionamentos com todas as opções para o nome Eldorado, é possível determinar a versão correta, e consequentemente, a desambiguação de nomes de lugares do texto. O método proposto para desambiguação de nomes de lugares trabalha com o apoio do OntoGazetteer e pode ser empregado na situação descrita. Esse processo é descrito a seguir.

4.3.1 Criação da coleção de notícias

A coleta de notícias para a criação dessa coleção foi feita em janeiro e fevereiro de 2011 (Tabela 4-4). Um coletor específico foi desenvolvido com o uso de XPath para extrair e armazenar o título e corpo da notícia para cada uma das fontes de notícias da Web como no estudo de caso anterior. Além disso, foram coletadas apenas notícias do estado de Minas Gerais. Para assegurar isso, as notícias foram obtidas a partir de seções locais ou seções estaduais das fontes de notícias.

[15] http://www.uai.com.br/htmls/app/noticia173/2010/07/22/noticia_minas,i=170175/index.shtml

Tabela 4-4 - Fontes de Notícias Web

Jornal Web	Coleta	Web site	Nome da seção do jornal	# Docs
Globominas	01 a 02/2011	http://globominas.globo.com/	Notícias gerais	90
O Tempo	01 a 02/2011	http://www.otempo.com.br/	Últimas notícias - Cidades	39
Uai	01 a 02/2011	http://www.uai.com.br/	Minas	21
Terra	01 a 02/2011	http://www.terra.com.br	Últimas notícias - Brasil	10
			Total	**160**

4.3.2 Desambiguação de nomes de lugares

O processo para desambiguação de nomes de lugares começa com os passos para detecção e inferência de nomes tratados no estudo de caso anterior. Sendo assim, após a coleta de notícias, um pré-processamento também foi realizado para retirar *stopwords* (ver Apêndice B-1). Em seguida, termos candidatos a nomes de lugares foram extraídos do texto da notícia, usando expressões regulares para identificar nomes próprios compostos ou não. Por fim, o reconhecimento e identificação do tipo dos nomes de lugares da notícia apoiado pelo OntoGazetteer, conforme descrito anteriormente.

Em seguida, o processo para desambiguação foi executado em uma sequência de passos. O Algoritmo 4-3 apresenta a função *Desambiguar*, que é responsável pela desambiguação de termos geográficos ambíguos. Dessa forma, o primeiro passo começa com a identificação de nomes ambíguos e não-ambíguos (isto é, nomes que ocorrem mais de uma vez ou apenas uma vez no *gazetteer* ontológico). Nesse estudo foi considerada apenas a ambigüidade *geo/geo*, descrita no Capítulo 2.

O segundo passo para desambiguação é a identificação de indicadores de localidade, ou seja, nomes de tipos de lugar no texto que geralmente antecedem o nome do lugar. Por exemplo, na Figura 4-7, a palavra *Bairro* aparece antes de Eldorado, e no OntoGazetteer o nome “Bairro Eldorado” existe como um nome alternativo para dois lugares diferentes. Portanto, se após esse passo ainda existe mais de um lugar associado a um nome, o processo de desambiguação deve prosseguir.

```
Função Desambiguar()
    {
      Para cada (Noticia listaDeNoticias)
       {
         Para cada (Frase na listaDeFrases)
           {
             Se (contemLugares)
              {
                listaDeTermosAmbiguos = Função RecuperarTermoGeoAmbiguo;
                Para cada (termoAmbiguo na listaDeTermosAmbiguos)
                 {
                   Se (listaIndicadores > 0)
                     {
                       Para cada(Indicador na listaIndicadores)
                         {
                           Para cada(Posicao na listaPosicao)
                             {
                               termoMaisProximo = Função recuperarTermoMaisProximo;
                               Para cada(Lugar na listaDeLugares)
                                {
                                  Função desambiguarPorIndicadoresDeLocalidade;
                                }
                             }//fim do Para cada Posição
                         }//Fim Para cada Termo auxiliar
                     }

                   Se (listaIndicadores == 0 || StatusLugar != 'DESAMBIGUADO')
                     {
                       ListaDeRelacionamentos = Função RecuperarRelacionamentosLugares;
                       Função desambiguarPorRelacionamento;
                     }
                 }
              } // fim Se contémLugar
           }//Fim Para cada Frase
       }//Fim Para cada Noticia
    }
```

Algoritmo 4-3 – Função Desambiguar

O próximo passo para desambiguação usa informações sobre os relacionamentos entre lugares do OntoGazetteer. Todas as relações entre os nomes de lugares candidatos, tanto implícitos como explícitos, são recuperados, armazenados e usados para a construção de um grafo de relacionamento entre os lugares, sendo que os lugares são os nós e os relacionamentos são os vértices. Um determinado nome do lugar será escolhido como o correto se for aquele com o maior número de relações com os outros lugares. A Figura 4-8 mostra o grafo construído com os nomes de lugar do texto da Figura 4-6. O subgrafo na Figura 4-8 (a) contém a versão mais provável para o nome de lugar Eldorado, uma vez que este nome é o que tem mais relacionamentos com outros nomes encontrados no texto.

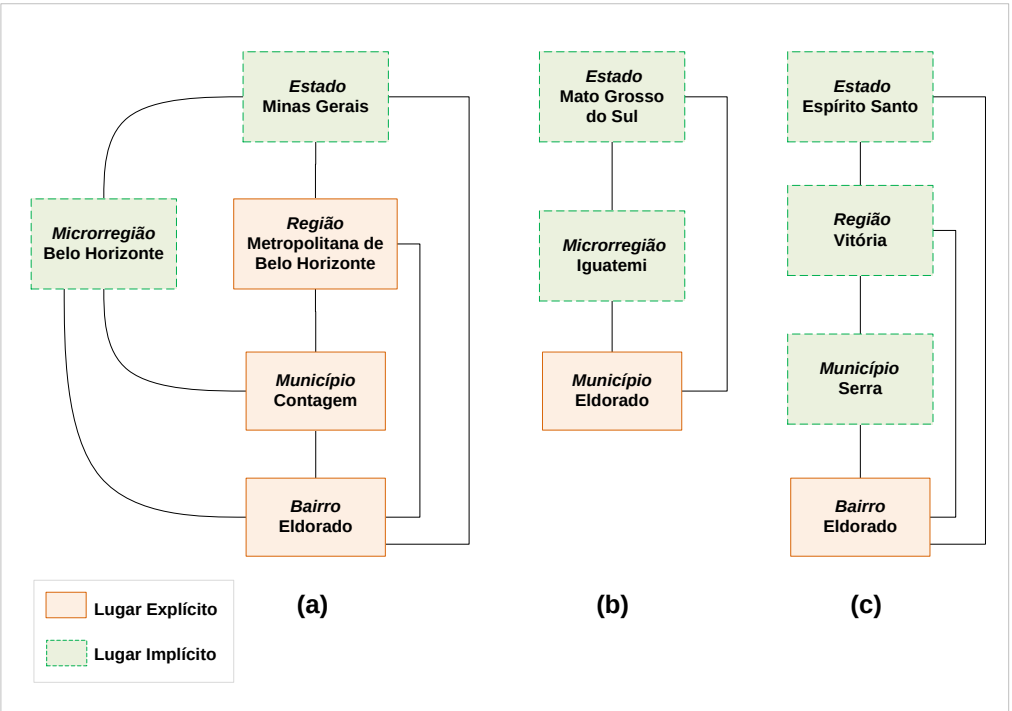
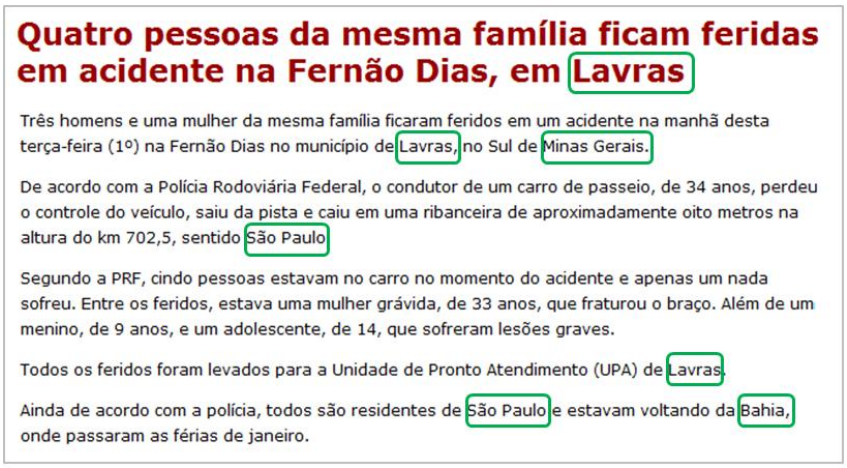


Figura 4-8 - Grafo de relacionamento

Um passo adicional para a desambiguação, que não foi explorado neste estudo de caso, consistiria em usar os vários tipos de relacionamentos armazenados no OntoGazetteeer. Por exemplo, podem ser levados em consideração apenas os relacionamentos hierarquicos (subdivisão territorial) e descartar o resto, ou considerar também as relações de vizinhança.

O processo de desambiguação apresentado depende da qualidade e abrangência do conteúdo do *gazetteer*. A fim de ilustrar um efeito negativo devido à falta de dados, considere a notícia apresentada na Figura 4-9. O contexto deste artigo é ambíguo, já que há referências a lugares aparentemente não relacionados. Por exemplo, considerando o conteúdo atual do OntoGazetteer, São Paulo é um estado brasileiro, uma cidade e um bairro de Belo Horizonte, Minas Gerais. A ligação potencial da última versão da referência com Minas Gerais pode levar a interpretações erradas, pois reforça uma conexão que não é a correta no contexto da notícia. Esse problema é causado pelo fato de que o nome Fernão Dias, que se refere a uma rodovia que liga Minas Gerais ao estado de São Paulo, passa pela cidade de Lavras, e a rodovia é o verdadeiro assunto do texto. No entanto, neste exemplo, Fernão Dias não está incluído no OntoGazetteer,

levando a um possível erro. Esse tipo de erro será menos frequente na medida que o conteúdo do OntoGazetteer expandir.



Quatro pessoas da mesma família ficam feridas em acidente na Fernão Dias, em Lavras

Três homens e uma mulher da mesma família ficaram feridos em um acidente na manhã desta terça-feira (1º) na Fernão Dias no município de Lavras, no Sul de Minas Gerais.

De acordo com a Polícia Rodoviária Federal, o condutor de um carro de passeio, de 34 anos, perdeu o controle do veículo, saiu da pista e caiu em uma ribanceira de aproximadamente oito metros na altura do km 702,5, sentido São Paulo.

Segundo a PRF, cindo pessoas estavam no carro no momento do acidente e apenas um nada sofreu. Entre os feridos, estava uma mulher grávida, de 33 anos, que fraturou o braço. Além de um menino, de 9 anos, e um adolescente, de 14, que sofreram lesões graves.

Todos os feridos foram levados para a Unidade de Pronto Atendimento (UPA) de Lavras.

Ainda de acordo com a polícia, todos são residentes de São Paulo e estavam voltando da Bahia, onde passaram as férias de janeiro.

Figura 4-9 - Exemplo de Notícia (fonte: OTempo, 01/02/2011)

Outra limitação do processo apresentado é o fato de que ele atualmente é limitado à ambigüidade *geo/geo*. Por exemplo, a palavra *Vitória* pode aparecer em um texto e ser claramente identificada como a capital do estado do Espírito Santo. No entanto, existe um time de futebol com exatamente este nome, mas está relacionado a Salvador, capital do estado da Bahia. Dependendo da ocorrência ou não de nomes de outros lugares no texto que indiquem uma referência para algum lugar no Espírito Santo, um erro de identificação pode ocorrer. Pretendemos melhorar isso expandindo a lista de termos relacionados aos lugares, já incluídos na estrutura do OntoGazetteer. Uma fonte viável de termos é a Wikipédia [Alencar et al., 2010].

4.3.3 Avaliação e resultados dos experimentos

Após executar os processos de reconhecimento e desambiguação de nomes de lugares na coleção de notícias coletadas, a avaliação do processo foi feita através de uma verificação manual da taxa de sucesso. Para facilitar essa verificação manual uma interface Web foi criada (Figura 4-10). Essa interface mostra o título e o corpo da notícia, e relaciona os nomes identificados automaticamente pelo nosso algoritmo. Para

cada um desses nomes, um voluntário deveria indicar se a desambiguação foi feita corretamente ou não. A Figura 4-10 apresenta a tela para avaliação da notícia dada no exemplo da Figura 4-7. Como já descrito o nome de lugar *Eldorado* é ambíguo, no entanto, pelo processo de desambiguação encontrou o tipo correto para o lugar.

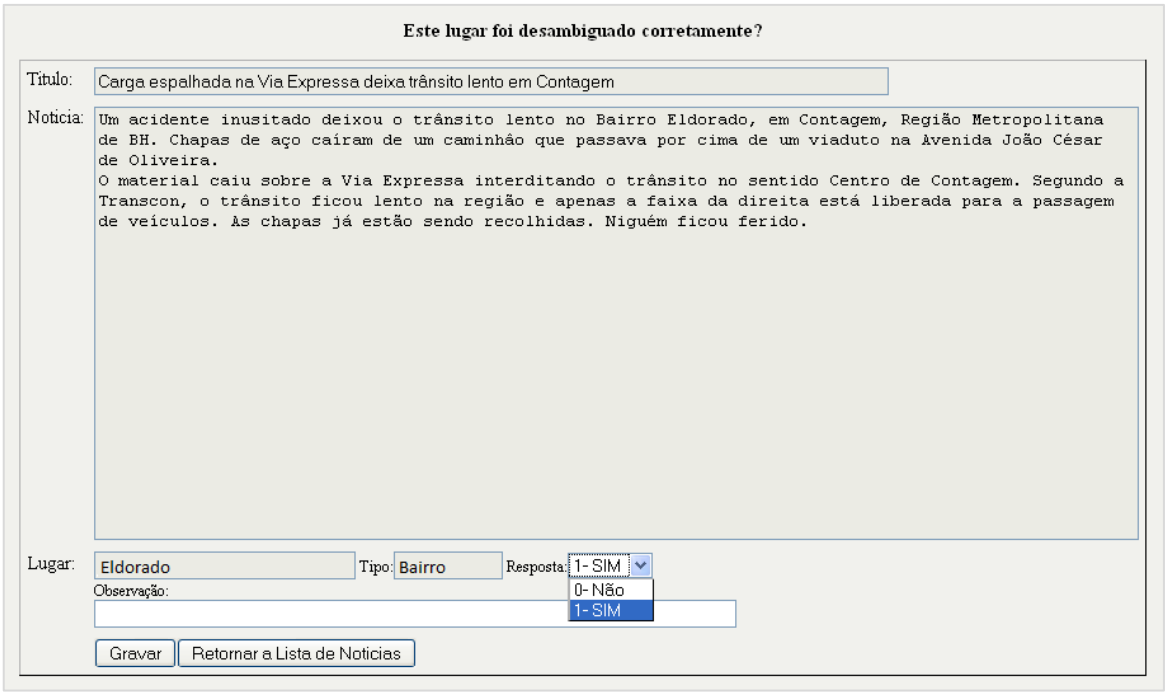


Figura 4-10 - Interface Web para avaliação dos voluntários

A coleção inicialmente contava com 160 notícias, das quais 152 notícias tiveram nomes de lugares identificados e dessas, 128 notícias tinham nomes de lugar ambíguos. Em média, cada notícia tinha quatro nomes de lugares, dos quais três eram ambíguos. Foram verificados manualmente 100% dos nomes de lugar desambiguados nas 128 noticias. Os resultados mostram que 346 nomes de lugares foram analisados, e 82% foram avaliados como desambiguados corretamente (Tabela 4-5).

Como mencionado anteriormente, a desambiguação foi executada em duas etapas consecutivas, uma etapa que usa o nome do tipo do lugar, também chamado indicador de localidade, se disponível, e a outra etapa que analisa o grafo de relacionamentos entre os lugares. A Tabela 4-6 mostra a proporção de desambiguações corretas e incorretas para cada etapa. Observe que o primeiro processo foi capaz de resolver

menos de 20% das ambigüidades, mas foi muito eficiente, com apenas um resultado incorreto. A incerteza é muito maior na segunda etapa; no entanto, já foi demonstrado que o acerto dessa etapa vai depender da qualidade e expansão do conteúdo do *gazetteer*.

Tabela 4-5 - Avaliação do número de nomes de lugares desambiguados

Desambiguado	#	%
Incorretamente	63	18,2
Corretamente	283	81,9
Total	**346**	**100,0**

Tabela 4-6 - Avaliação do número de lugares desambiguados de acordo com o processo

Processo de Desambiguação	Avaliação	#	%
1- Através do indicador de localidade	Incorreta	1	0,3
	Correta	63	18,2
2- Através do relacionamento entre os lugares	Incorreta	62	17,9
	Correta	220	63,6
Total		**346**	**100,0**

A Figura 4-11 e a Tabela 4-7 mostram o percentual de desambiguação correta e incorreta de acordo com o número de lugares mencionados no texto. É observada uma pequena variação na taxa de sucesso com o aumento do número de lugares. Seria razoável esperar que os casos em que há poucos ou muitos nomes no texto gerassem uma taxa menor de sucesso. Por exemplo, com apenas um nome de lugar ambíguo, em qualquer texto, não há informações suficiente para ajudar na desambiguação. Por outro lado, se há nomes demais, aumentam as chances de confusão na análise do grafo e também as chances de erro. No entanto, nos nossos experimentos, não foi esse o caso: a taxa de erro é um pouco maior quando existem mais nomes de lugar no texto, mas a diferença para as outras situações é pequena.

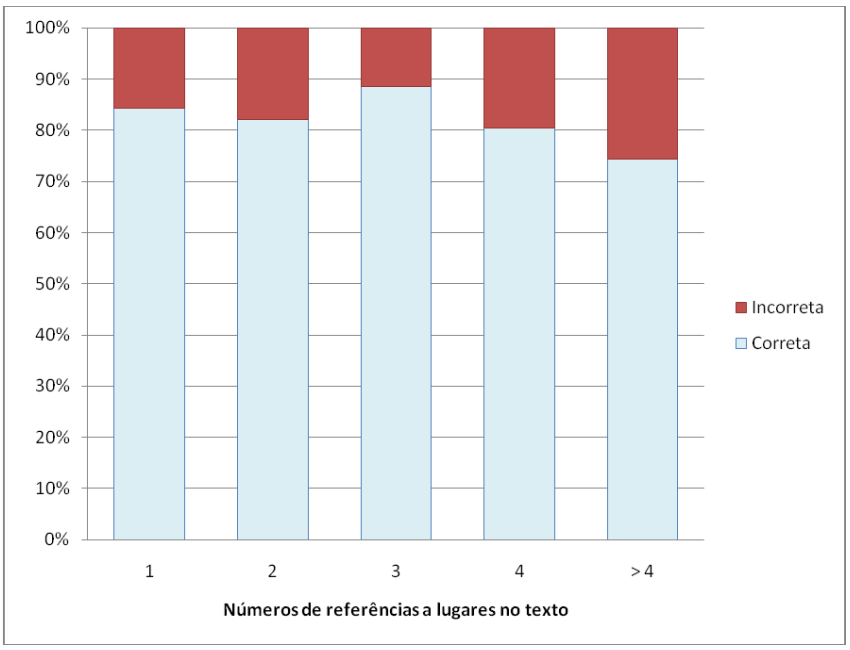


Figura 4-11 – Gráfico da eficiência segundo o número de lugares mencionados no texto

Esses experimentos mostraram que o processo de desambiguação proposto é promissor, tendo alcançado bons resultados, mesmo considerando que o conteúdo do OntoGazetteer ainda está sendo expandido. Sendo assim, o processo proposto pode melhorar com as listas de termos relacionados aos lugares, que estão em desenvolvimento e devem cobrir algumas ambigüidades *geo/não-geo*.

Tabela 4-7 - Eficiência de acordo com o número de lugares mencionados no texto

Número de referências para lugares no texto	# Correta	# Incorreta
1 lugar	16	3
2 lugares	87	19
3 lugares	77	10
4 lugares	45	11
Mais de 4 lugares	58	20
Total	**283**	**63**

4.4 Conclusão

Em suma, a estrutura do OntoGazetteer suportou a demanda dos estudos de caso, porém foram detectadas necessidades quanto à expansão do seu conteúdo, que representará mais abrangência e mais qualidade em tarefas de recuperação de informação geográfica. Como os dois estudos de casos possuíam um escopo pequeno foi possível avaliar e validar a estrutura proposta para o *gazetteer*, bem como apresentar bons resultados para os experimentos. O primeiro estudo de caso tinha como objetivo detectar nomes de lugares e inferir contexto geográfico a partir do relacionamento entre os nomes de lugares explícitos identificados nos textos. O segundo estudo de caso tinha como objetivo desambiguar nomes de lugares de notícias a partir de indicadores de localidade e através dos relacionamentos entre os lugares. Assim como o primeiro, o segundo estudo de caso começa com a detecção dos nomes de lugares com o apoio do OntoGazetteer. Dessa forma, quanto melhor a base de dados do *gazetteer* melhor serão os resultados para detectar os nomes de lugares no texto e, consequentemente, melhor será o resultado para inferir contexto geográfico e desambiguar nomes de lugares.

Capítulo 5

Conclusão e Trabalhos Futuros

O trabalho apresentou uma nova estrutura para os *gazetteers*, dicionários de nomes de lugares, que visa diminuir as suas atuais limitações e permite apoiar tarefas de recuperação de informação geográfica. O *gazetteer* ontológico ou OntoGazetteer utiliza conceitos de ontologia para definir uma forma mais flexível de estabelecer e manter relacionamentos semanticamente mais ricos entre os lugares, além de adicionar componentes para manter nomes alternativos e listas de termos relacionados aos lugares. Os relacionamentos podem ser geográficos ou topológicos, além de permitir que sejam criadas conexões semânticas entre lugares geograficamente distintos.

O trabalho também descreve algumas maneiras como o *gazetteer* proposto pode ser usado em tarefas de recuperação de informação geográfica, além de apresentar estudos de caso em duas dessas tarefas para validar a estrutura proposta para o *gazetteer* ontológico.

O primeiro estudo de caso apresentou uma estratégia para detecção e inferência de contexto geográfico em notícias da Web com o uso do OntoGazetteer. O estudo mostrou que o uso do *gazetteer* ontológico é muito valioso, já que através dos relacionamentos armazenados é possível inferir lugares que não foram explicitamente mencionados no texto. O reconhecimento de nomes de lugares teve bons resultados principalmente para os tipos de lugares mais conhecidos e que são mais comumente encontrados em notícias, como municípios, bairros, rodovias e estados.

O segundo estudo de caso utilizou a estratégia anterior para detecção e inferência

de contexto geográfico, além de propor uma estratégia para desambiguação de nomes de lugares em notícias da Web através de duas etapas. A primeira etapa utiliza indicadores de localidade que antecedem os nomes de lugares. Caso essa etapa não consiga desambiguar o nome, uma segunda etapa é utilizada. A segunda etapa cria um grafo de relacionamento entre os nomes de lugares reconhecidos tanto explicitamente como implicitamente no texto. A partir do grafo de relacionamento, é possível desambiguar os nomes de lugares ao considerar somente as conexões mais fortes com os demais nomes mencionados no texto. A estratégia proposta para desambiguar nomes de lugares obteve bons resultados. No entanto, mostrou que pode melhorar muito com a expansão do conteúdo do OntoGazetteer. Por exemplo, informações sobre bacias hidrográficas e rodovias podem servir como fonte para traçar relacionamentos entre grupos de lugares, por exemplo, duas cidades não vizinhas podem estar relacionadas pelo fato de ambas se situarem ao longo da mesma rodovia ou na mesma bacia hidrográfica

5.1 Trabalhos futuros

À medida que o OntoGazetteer tem seu conteúdo melhorado e expandido, o seu uso nas tarefas descritas e em tantas outras torna-se mais abrangente e com mais qualidade. Portanto, o principal trabalho futuro, e que será um processo constante, é a expansão do conteúdo do *gazetteer* ontológico, tanto quanto possível utilizando informações de diversos bancos de dados geográficos com dados do Brasil como um todo. Para isso, serão utilizadas fontes livres, mas também contatos com prefeituras e outros órgãos que tenham esses dados. A expansão para lugares internacionais não é inconcebível, embora seja algo trabalhoso.

A partir dos dados geográficos carregados no banco de dados geográfico é possível derivar facilmente os relacionamentos geográficos e topológicos. Relacionamentos semânticos também estão sendo expandidos considerando inicialmente relacionamentos geográficos indiretos, como por exemplo, dois municípios por onde um mesmo rio passa estão relacionados por este motivo, mesmo que esses municípios são sejam adjacentes.

Outro foco para trabalhos futuros e também um trabalho em paralelo é encontrar

as listas de termos relacionados aos diversos lugares armazenados no OntoGazetteer. Será usada como base de conhecimento a Wikipédia, que já tem mostrado resultados promissores [Alencar et al., 2010]. Além de fontes de dados de colaboração voluntária para *geotagging*, como por exemplo, Wikimapia.

Com relação aos estudos de caso, um trabalho futuro é melhorar o algoritmo para desambiguação de nomes de lugares com a utilização de pesos diferentes para tipos de relacionamentos diferentes entre os lugares para a construção do grafo de relacionamentos, bem como a utilização de outros tipos de relacionamentos e não apenas o hierárquico.

Por fim, é necessário melhorar usabilidade da interface Web, que é o mecanismo pelo qual o usuário terá acesso aos dados do OntoGazetteer de forma segura. Pretende-se criar novos serviços Web em busca de atender melhor as necessidades de sistemas de GIR e minimizar o acesso aos dados geográficos. Além de permitir que voluntários possam contribuir com bases de dados de diversas regiões do Brasil. Para isso, será incluído na interface uma área no qual o usuário poderá enviar bases de dados para serem analisadas e posteriormente, carregadas no OntoGazetteer.

Referências bibliográficas

Adriani, M. & Paramita, M. L. (2007). Identifying location in indonesian documents for geographic information retrieval, *Proceedings of the 4th workshop on GIR*. Lisbon, Portugal: ACM.

Alencar, R. O., Davis-Jr., C. A. & Gonçalves, M. A. (2010). Geographical classification of documents using evidence from Wikipedia, *Proceedings of the 6th Workshop on GIR*. Zurich, Switzerland: ACM.

Amitay, E., Har'El, N., Sivan, R. & Soffer, A. (2004). Web-a-where: geotagging web content, *Proceedings of the 27th annual international SIGIR Conference on Research and Development in Information Retrieval*. Sheffield, United Kingdom: ACM.

Backstrom, L., Kleinberg, J., Kumar, R. & Novak, J. (2008). Spatial variation in search engine queries, *Proceeding of the 17th International Conference on WWW*. Beijing, China: ACM.

Baeza–Yates, R. & Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. New York.

Barbosa, I. & Casanova, M. A. (2010). *Dicionário de tipos de feições geoespaciais para o território brasileiro* Paper presented at the Proceedings of the XI Brazilian Symposium on Geoinformatics, Campos do Jordão, SP, Brazil.

Borges, K. A. V. (2006). *Uso de uma Ontologia de Lugar Urbano para Reconhecimento e Extração de Evidências Geo-espaciais na* UFMG, Belo Horizonte.

Borges, K. A. V., Davis-Jr, C. A. & Laender, A. H. F. (2001). OMT-G: An object-oriented data model for geographic applications. *GeoInformatica, 5*(3), 221-260.

Borges, K. A. V., Laender, A. H. F., Medeiros, C. B. & Davis Jr, C. A. (2007). Discovering geographic locations in web pages using urban addresses, *Proceedings of the 4th workshop on GIR*. Lisbon, Portugal: ACM.

Brisaboa, N. R., Luaces, M. R., Places, Á. S. & Seco, D. (2010). Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index. *Geoinformatica, 14*(3), 307-331.

Buscaldi, D. & Rosso, P. (2007). A comparison of methods for the automatic identification of locations in wikipedia, *Proceedings of the 4th ACM workshop on Geographical information retrieval*. Lisbon, Portugal: ACM.

Chaves, M. S. (2009). *Uma Metodologia para Construção de Geo-Ontologias.* Universidade de Lisboa, Lisboa.

Davis-Jr, C. A. & Laender, A. H. F. (1999). Multiple representations in GIS: materialization through map generalization, geometric, and spatial analysis operations, *Proceedings of the 7th ACM international symposium on Advances in geographic information systems*. Kansas City, Missouri, United States: ACM.

Delboni, T. M., Borges, K. A., Laender, A. H. & Davis-Jr, C. A. (2007). Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS, 3*, 377-397.

Fu, G., Jones, C. B. & Abdelmoty, A. I. (2005). Ontology-Based Spatial Query Expansion in Information Retrieval. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE* (Vol. 3761, pp. 1466-1482): Springer Berlin / Heidelberg.

Goodchild, M. F. & Hill, L. L. (2008). Introduction to digital gazetteer research. *Int. J. Geogr. Inf. Sci., 22*(10), 1039-1044.

Gouvêa, C. (2009). *Uma Abordagem para o Enriquecimento de Gazetteers a partir de Notícias visando o Georreferenciamento de Textos na Web.* Universidade Católica de Pelotas, Pelotas.

Gouvêa, C., Loh, S., Garcia, L. F. F., Fonseca, E. B. & Wendt, I. (2008). *Discovering Location Indicators of Toponyms from News to Improve Gazetteer-Based Geo-Referencing* Paper presented at the Simpósio Brasileiro de Geoinformática - GEOINFO Rio de Janeiro, RJ.

Hill, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints, *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*: Springer-Verlag.

Janowicz, K. & Kessler, C. (2008). The role of ontology in improving gazetteer interaction. *Int. J. Geogr. Inf. Sci., 22*(10), 1129-1157.

Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., Kreveld, M. v., *et al.* (2002). Spatial information retrieval and geographical ontologies an overview of the SPIRIT project, *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. Tampere, Finland: ACM.

Jones, C. B. & Purves, R. S. (2008). Geographical information retrieval. *Int. J. Geogr. Inf. Sci., 22*(3), 219-228.

Leidner, J. L. (2004). Towards a reference corpus for automatic toponym resolution evaluation, *SIGIR*. Sheffield, UK (Geographic IR Workshop).

Leidner, J. L. (2007). *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* University of Edinburgh, Scotland, UK.

Lopez-Pellicer, F. J., Silva, M. J. & Chaves, M. (2010). *Linkable Geographic Ontologies.* Paper presented at the GIR´10, Zurich, Switzerland.

Machado, I. M. R., Alencar, R. O., Campos-Jr, R. O. & Davis-Jr, C. A. (2010). *An Ontological Gazetteer for Geographic Information Retrieval* Paper presented at the Proceedings of the XI Brazilian Symposium on Geoinformatics, Campos do Jordão, SP, Brazil.

Overell, S. & Ruger, S. (2006). *Identifying and grounding descriptions of places.* Paper presented at the SIGIR Workshop on GIR, Seattle.

Overell, S. & Ruger, S. (2008). Using co-occurrence models for placename disambiguation. *Int. J. Geogr. Inf. Sci., 22*(3), 265-287.

Overell, S. E. & Ruger, S. (2007). Geographic co-occurrence as a tool for gir, *Proceedings of the 4th ACM workshop on GIR*. Lisbon, Portugal: ACM.

Popescu, A., Grefenstette, G. & Moellic, P. A. (2008). Gazetiki: automatic creation of a geographical gazetteer, *Proceedings of the 8th ACM/IEEE-CS JCDL*. Pittsburgh, PA, USA: ACM.

Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T., Blackler, K., *et al.* (2006). *Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation.* Paper presented at the Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006).

Sanderson, M. & Kohler, J. (2004). *Analyzing geographic queries.* Paper presented at the Workshop on Geographic Information Retrieval (GIR 2004), Sheffield, UK.

Silva, M. J., Martins, B., Chaves, M., Afonso, A. P. & Cardoso, N. (2004). *Adding Geographic Scopes to Web Resources.* Paper presented at the SIGIR'04 - Workshop on GIR, Sheffield, UK.

Smith, D. A. & Crane, G. (2001). Disambiguating Geographic Names in a Historical Digital Library, *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*: Springer-Verlag.

Souza, L. A., Davis-Jr, C. A., Borges, K. A. V., Delboni, T. M. & Laender, A. H. F. (2005). The Role of Gazetteers in Geographic Knowledge Discovery on the Web, *Proceedings of the 3th LA-Web Congress*: IEEE Computer Society.

Souza, L. A., Delboni, T. M., Borges, K. A. V., Davis-Jr., C. A. & Laender, A. H. F. (2004). *Locus: Um Localizador Espacial Urbano.* Paper presented at the VI Brazilian Symposium on Geoinformatics, 22-24, Campos do Jordão, São Paulo, Brazil.

Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H. & Sperling, J. (2008). NewsStand: a new view on news, *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. Irvine, California: ACM.

Toral, A. & Munoz, R. (2006). *A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia.* Paper presented at the EACL 2006.

Uryupina, O. (2003). Semi-supervised learning of geographical gazetteers from the internet, *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Vol 1*: Association for Computational Linguistics.

Volz, R., Kleb, J. & Mueller, W. (2007). Towards ontology-based disambiguation of

geographical identifiers, *WWW2007*. Banff, Canada.

Wang, C., Xie, X., Wang, L., Lu, Y. & Ma, W.-Y. (2005). Web resource geographic location classification and detection, *Special interest tracks and posters of the 14th international conference on WWW*. Chiba, Japan: ACM.

Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.-Y., *et al.* (2005). Detecting dominant locations from search queries, *Proceedings of the 28th annual international SIGIR conference on Research and Development in Information Retrieval*. Salvador, Brazil: ACM.

Zong, W., Wu, D., Sun, A., Lim, E.-P. & Goh, D. H.-L. (2005). On assigning place names to geography related web pages, *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*. Denver, CO, USA: ACM.

Apêndice A

A.1 Script para carga de dados da tabela de relacionamento geográfico (Relacionamento Hierárquico)

```
<?php

$db = "host='localhost' port='5432' dbname='****' user='postgres'
password='*****'";
$bd = pg_connect($db);


// nome do arquivo
$arquivo = '//relgeo_hierarquico_bairro-muni.csv';

// ponteiro para o arquivo
$fp = fopen($arquivo, "r");
$data = date(Y)."-".date(m)."-".date(d)."";
// processa os dados do arquivo
while(($dados = fgetcsv($fp, 0, ";")) !== FALSE)
     {

       $query1 = "INSERT INTO \"relacionamento_geo\" (codlugar1,
codlugar2, codtiporelgeo, datacriacao)
       VALUES ($dados[0], $dados[1], $dados[2]    , '$data'    );";
       $query2 = "SELECT * FROM \"relacionamento_geo\" WHERE
codtiporelgeo = $dados[2] and codlugar1 = $dados[0] and codlugar2
= $dados[1]";
       $res2 = pg_query($query2);
       $numlinhas = pg_num_rows($res2);
       echo "Resultado: $numlinhas ";

       if ($numlinhas == 0)
           {
             $res1 = pg_query($query1);
           }

      echo "$dados[0];$dados[1];$dados[2];$data";
      echo "<br>";

     }

fclose($fp);
?>
```

A.2 Script para carga de dados da tabela de relacionamento geográfico (Relacionamento Vizinhança)

```
public class GeraVizinhos {
     static Connection connection;
     public static void insertRelGeo(ResultSet result) throws
SQLException {
          try {
               while (result.next()) {
                    String query = "INSERT INTO
relacionamento_geo(codlugar1, codlugar2, codtiporelgeo,
datacriacao) VALUES (" + result.getString(1)
                      + ","
                      + result.getString(2)
                      + ","
                      + result.getString(3) + ",'" + data() +
"');";
Statement statement = connection.createStatement();
     statement.executeUpdate(query);
                    }
               }
          }
public static ResultSet retrieveRelacionamentoGeo(int i)
               throws SQLException {
               String query = "select DISTINCT l1.codlugar AS
lugar1, l2.codlugar AS lugar2, rel.codtiporelgeo AS
relacionamento " + "from lugar AS l1, lugar AS l2, area AS
ga1, area AS ga2, tipo_relacionamento_geo AS rel " + "WHERE
l1.codtipolugar = " + i + " AND l2.codtipolugar = " + i
          + " AND rel.codtipolugar1 = l1.codtipolugar AND
rel.codtipolugar2 = l2.codtipolugar " + " AND ga1.codlugar =
l1.codlugar AND ga2.codlugar = l2.codlugar AND
ST_TOUCHES(ga2.geom, ga1.geom) ORDER BY lugar1 ASC";
          Statement statement = connection.createStatement();
          ResultSet places = statement.executeQuery(query);
          return places;
```

```
     }
public static String data() {
          Date hoje = new Date(System.currentTimeMillis());
          String formato = "yyyy-MM-dd";
          SimpleDateFormat        formatada        =        new
SimpleDateFormat(formato);
          return formatada.format(hoje);
        }

public static void main(String[] args) {
          "jdbc:postgresql:****";
          String url = "****";
          String username = "postgres";
          String userpass = "*****";
          String driver = "org.postgresql.Driver";
          int tipoDeLugares[] = { };
     try {
               Class.forName(driver);
               connection             =             (Connection)
DriverManager.getConnection(url,
               username, userpass);
               System.out.println("Connection done!");
               for (int i = 0; i < tipoDeLugares.length; i++) {
                    ResultSet             places             =
retrieveRelacionamentoGeo(tipoDeLugares[i]);
                    insertRelGeo(places);
               }
               connection.close();
               System.out.println("Fechou conexão");
               System.exit(0);
        }
}
```

Apêndice B

B.1 Lista de *Stopwords*

Para, para, PARA
Com, com, COM
Ao(s), ao(s), AO(S)
Em, em, EM
O(s), O(S)
A(s), A(S)
No(s), no(s), NO(S)
Na(s), na, NA(S)
Além
Um, UM, Uns, UNS
Uma(s), UMA(S)
Se, SE