TIAGO MARQUES DELBONI

EXPRESSÕES DE POSICIONAMENTO COMO FONTE DE CONTEXTO GEOGRÁFICO NA *WEB*

Belo Horizonte
26 de agosto de 2005

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

EXPRESSÕES DE POSICIONAMENTO COMO FONTE DE CONTEXTO GEOGRÁFICO NA *WEB*

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

TIAGO MARQUES DELBONI

Belo Horizonte
26 de agosto de 2005

Resumo

Expressões de posicionamento são expressões em linguagem natural que descrevem a posição de um objeto de interesse em relação a um ponto de referência, como, por exemplo, 'a duas quadras da praia de Ipanema'. Ao contrário de outros localizadores geográficos, pouco se sabe a respeito das expressões de posicionamento. Este trabalho procura preencher essa lacuna mediante a realização de um estudo de caso envolvendo expressões de posicionamento na cidade de Belo Horizonte. Apesar de não serem tão precisas quanto um endereço exato, a correta interpretação das expressões de posicionamento permite–nos utilizá–las com sucesso em diversas aplicações de recuperação de informação geográfica. Quando utilizadas em um experimento de busca local, por exemplo, as expressões de posicionamento obtiveram respostas com uma precisão média cerca de 60% superior à obtida pelo *Google*.

Abstract

Positioning expressions are natural language expressions that describe the position of a subject of interest relatively to a landmark, for example, 'two blocks from the Ipanema beach'. In contrast to other geographic localizers, little is known regarding the positioning expressions. This work aims to fill this gap presenting a case study involving positioning expressions in the city of Belo Horizonte. Although not so accurate as an address, the correct interpretation of positioning expressions allows them to be successfully employed in diverse geographic information retrieval applications. When used in a local search experiment, for example, the average precision of the answers obtained using positioning expressions was 60% higher then that obtained by *Google*.

Aos meus Pais

"Moe... Gimme a beer!"
Homer Simpson

Agradecimentos

Queridos pais, Marco e Zélia, muito obrigado pelo amor, paciência, compreensão, apoio e incentivo recebidos durante toda a jornada. Se aqui cheguei, foi graças a vocês!

Alberto, certamente tenho muito a te agradecer. Obrigado pela confiança, oportunidades e ensinamentos. Foi uma grande honra tê–lo como orientador. Esteja certo de que apreendi muito.

Karla, você não existe! Sem sua ajuda, incentivo e companheirismo as coisas seriam bem mais difíceis.

Agradeço também ao professor Berthier pela oportunidade de ingressar na área de pesquisa, ao pessoal do SIDS pelo apoio irrestrito, ao Palmieri pelas sábias palavras de conforto e incentivo — "nada é tão ruim que não possa piorar" — e ao Luís pelos inúmeros exemplos práticos fornecidos.

Obrigado aos colegas e amigos que me ajudaram nessa empreitada e fizeram-na valer a pena. Valeu pessoal!

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Segundo Himmelstein (2005), as tendências que apontam para o rápido crescimento da utilização da Internet como fonte de informações relacionadas a um escopo geográfico são inconfundíveis. A penetração da Internet na vida das pessoas, com a disseminação do acesso à rede, especialmente em banda–larga, fez aumentar a intensidade de utilização dessa mídia como ferramenta de auxílio a atividades do dia–a–dia como encontrar locais, produtos, lojas, serviços e eventos na vizinhança ou cidade onde moram ou trabalham. A Internet ocupa agora um espaço que até então era privilégio de mídias tradicionais como jornais locais, classificados, páginas amarelas e catálogos de produtos. Uma pesquisa de janeiro de 2004, realizada pelo *The Kelsey Group* e pelo *site BizRate.com*, revelou que 25% das buscas de compradores *on–line* eram por produtos e serviços "próximos à minha casa ou trabalho".[1]

Recorrer a uma máquina de busca para encontrar informações de interesse na Internet é tarefa rotineira para a maioria das pessoas. Cruciais para o desenvolvimento da Internet como a conhecemos hoje e para a difusão sem fronteiras do conhecimento, as máquinas de busca nos permitem encontrar, com rapidez e boa precisão, informações relevantes para uma ampla gama de consultas em meio ao universo em contínua expansão dos bilhões de documentos publicados na *Web*.

No entanto, a utilização de máquinas de busca para a realização de pesquisas por critérios geográficos, conhecida como busca local (*local search*), apresenta uma série de limitações. Essas limitações devem–se não apenas à maneira como os dados da *Web* estão estruturados, mas também aos métodos atuais de recuperação de informação, tipicamente limitados a casamentos exatos ou parciais entre um conjunto de palavras–chave, denominado *consulta*, e o texto dos documentos. Por exemplo, a inclusão do nome de um lugar entre as palavras–chave não assegura que apenas documentos contendo referências a esse lugar sejam recuperados. Pelo contrário, todos os documentos

[1] `http://www.kelseygroup.com/press/pr040211.htm`

contendo as palavras–chave da consulta serão recuperados, mesmo que no contexto de alguns deles o nome especificado não se refira a um lugar. Por exemplo, a palavra 'Vitória' pode, além de uma cidade, referir–se a um triunfo, uma pessoa ou ainda a uma família de plantas aquáticas. Em outras palavras, uma vez que a informação disponível é acessada de maneira uniforme, as máquinas de busca tradicionais ignoram o contexto geográfico dos documentos da *Web*, incluindo na resposta documentos geograficamente irrelevantes (Ding et al., 2000; McCurley, 2001; Amitay et al., 2004; Silva et al., 2004).

Para tornarem–se aptas ao tratamento de consultas mais específicas, ligadas a domínios especializados, exige–se das máquinas de busca um nível superior de interpretação semântica da necessidade de informação do usuário, interfaces mais elaboradas e amigáveis para a especificação da consulta e a visualização dos resultados, e estratégias diferenciadas para seleção e ordenação dos documentos relevantes. Essas novas estratégias devem levar em consideração evidências adicionais além daquelas utilizadas atualmente, isto é, a freqüência de palavras–chave no documento e na coleção e a existência de *links* entre os documentos.

Acompanhando essa tendência, a maioria dos grandes portais e máquinas de busca, como *AOL*, *MSN*, *Google*, *Yahoo!* e *Ask Jeeves*, passou a oferecer serviços voltados para a busca de conteúdo local, embora, em geral, disponíveis apenas para os Estados Unidos e Canadá. Alguns contemplam países da Europa, como o Reino Unido, e também o Japão. Uma deficiência presente em quase todas essas iniciativas é a falta de integração com dados provenientes da *Web* — a base de documentos onde as buscas geográficas são realizadas é tipicamente formada por registros cadastrais de empresas, adquiridos a partir da compra de bases comerciais de serviços tradicionais de páginas amarelas ou coletados de *sites* que mantêm catálogos desse tipo *on–line*, como *Citysearch.com* e *SuperPages.com*. Portanto, a *Web*, como importante repositório de informações geográficas que é (Borges et al., 2003), está sendo sub–utilizada pelas atuais ferramentas de busca local, que não exploram o potencial que ela tem a oferecer.

No domínio geográfico, a presença de elementos no texto de um documento, como endereços, códigos postais ou números de telefone, pode ser usada como evidência a partir da qual é possível correlacionar parte ou a totalidade das informações nele contidas a uma localização geográfica. De acordo com Himmelstein (2005), estima–se que pelo menos 20% dos documentos da *Web* incluam um ou mais desses elementos, fontes de evidência de contexto geográfico, denominados localizadores geográficos. Entre eles, as *expressões de posicionamento* descrevem, em linguagem natural, a posição de um objeto de interesse em relação a um ponto de referência, como, por exemplo, 'perto do Palácio das Artes', 'a duas quadras da praia de Ipanema' e 'no coração de São Paulo'.

Apesar de, sob o ponto de vista da geocodificação, não serem tão precisas quanto um endereço exato, carregando um certo grau de incerteza, as expressões de posicionamento

determinam uma relação no espaço que é mais fácil de ser compreendida e imaginada — a margem de erro introduzida por elas é compensada pela capacidade de raciocínio qualitativo e espacial que os seres humanos possuem. Dessa forma, acredita–se que a correta interpretação das expressões de posicionamento leve a uma geocodificação tão importante quanto a de um endereço. Inclusive, em algumas situações, as duas representações podem ser complementares.

Ao contrário de outras fontes de evidência de contexto geográfico, muito pouco se sabe a respeito das expressões de posicionamento, suas características, incidência, interpretações e aplicações. Estudar as expressões de posicionamento e identificar as situações em que elas possam ser empregadas no contexto de busca local, especialmente quando fazem referências a locais intra–urbanos, foram os objetivos que nortearam a execução deste trabalho, cujas principais contribuições são:

- Apresentar um estudo de caso que analisa expressões de posicionamento encontradas em documentos da *Web* e que estão relacionadas a locais na cidade de Belo Horizonte. Como resultado, foi possível determinar quais os tipos de local mais usados como ponto de referência e como local de interesse, as relações espaciais mais importantes, os tipos de objetos de interesse cuja posição é descrita por expressões de posicionamento e a presença dessas expressões em documentos da *Web*, entre outros (Seção 3.2);

- Desenvolver um método para identificar expressões de posicionamento em documentos textuais. Esse método baseia–se em regras de análise sintática de construções em linguagem natural e não requer um *gazetteer* para encontrar um nome e determinar se ele é ou não um local. Pelo contrário, um repositório de nomes de locais é obtido como resultado da análise dos documentos por um programa extrator implementado para esse fim. Cerca de 90% dos nomes desse repositório correspondem a nomes de locais (Seção 3.3);

- Determinar que, no contexto de uma aplicação de busca local, as principais relações espaciais em linguagem natural podem ser consideradas equivalentes em termos da distância que descrevem em relação a um ponto de referência intra–urbano. Esse fato foi explorado em uma aplicação baseada em expansão de consultas que utiliza o *Google* para processá-las (Delboni et al., 2005). As consultas expandidas apresentaram uma precisão média cerca de 60% superior à das consultas originais. Essa aplicação mostra ainda que os pontos de referência podem ser utilizados como centros de busca na especificação do escopo geográfico das consultas e que tanto a resolução de ambigüidade de nomes quanto a utilização de

coordenadas geográficas não são estritamente necessárias para se construir uma aplicação de busca local (Seções 3.4 e 4.1);

- Descrever uma estrutura de dados para armazenar referências geográficas, denominada grafo de inferência geográfica. Operando sobre essa estrutura é possível determinar a proximidade entre dois locais representados no grafo, mesmo que não estejam diretamente relacionados por meio de uma relação espacial. Outra operação possível, é calcular a coordenada geográfica aproximada para a posição de um local a partir das coordenadas de outros locais. Com essas informações, podemos aumentar a gama de aplicações e a eficácia dos métodos de recuperação de informação geográfica que utilizam informações provenientes das expressões de posicionamento (Seção 4.2).

Esta dissertação está organizada da seguinte forma. No Capítulo 2 há uma descrição dos serviços disponíveis para busca local na *Web* e uma revisão da literatura sobre o processo de atribuição de contexto geográfico a documentos textuais, especialmente a partir da identificação de nomes de lugares. Alguns conceitos básicos e terminologia são introduzidos no Capítulo 3, em que, através de uma série de experimentos, busca–se fazer uma caracterização das expressões de posicionamento. Um método para extrair expressões de posicionamento de documentos textuais também é apresentado. O Capítulo 4 discute aplicações onde as expressões de posicionamento podem ser utilizadas, apresentando alguns resultados práticos, e introduz o que denominamos grafo de inferência geográfica, uma estrutura de dados utilizada para derivar informações de cunho geográfico a partir das expressões de posicionamento. Por fim, as conclusões e desenvolvimentos futuros do trabalho são discutidos no Capítulo 5.

Capítulo 2

Recuperação de Informação Geográfica na Web

Segundo Larson (1996), a recuperação de informação geográfica (RIG) é uma área de pesquisa aplicada que combina aspectos de bancos de dados, interação homem–máquina, sistemas de informação geográficos (SIG) e recuperação de informação (RI). A maioria dos problemas e aplicações tradicionais de RI possuem uma versão equivalente em RIG, onde a dimensão geográfica associada às fontes de informação é o centro das atenções. Neste trabalho, a ênfase é no emprego das expressões de posicionamento como fonte de contexto geográfico para aplicações de busca local, caracterizadas na Seção 2.1. Na Seção 2.2 são discutidos alguns tópicos de interesse relacionados à identificação de expressões de posicionamento e atribuição de contexto geográfico proveniente dessas expressões.

2.1 Busca Local

O objetivo de uma aplicação de busca local é encontrar locais que ofereçam produtos e serviços de interesse do usuário e que estejam situados dentro de um escopo geográfico determinado por ele. 'Peças de Jeep em São Paulo' ou 'Serviço de impressão a laser próximo à Rua da Bahia, 2115 – Belo Horizonte' são consultas típicas desse tipo de aplicação, que poderia recuperar documentos a respeito de locais como revendedores de autopeças, oficinas, restauradores e colecionadores, para a primeira consulta, e gráficas, copiadoras ou editoras para a segunda.

Em uma aplicação de busca local, a consulta que descreve uma necessidade de informação é geralmente constituída por duas partes. A primeira, denominada *objeto de interesse*, consiste em um conjunto de palavras–chave representando o serviço ou produto que se espera encontrar nos locais retornados na resposta. A segunda determina

o *escopo geográfico*, isto é, a área dentro da qual os locais devem estar fisicamente instalados. O escopo geográfico é tipicamente formado por um *centro* e um *raio de busca*, que juntos definem uma área sobre a superfície terrestre. Localizadores associados a coordenadas geográficas, como endereços, códigos postais, nomes de lugares (especialmente nomes de cidade) ou as próprias coordenadas geográficas são utilizados como centro de busca. No caso do código postal e nome de lugar utiliza-se, por exemplo, a coordenada do centróide da área à qual eles correspondem. O raio de busca, por sua vez, é um valor representando a distância máxima do centro de busca onde um local pode estar situado e é, geralmente, escolhido a partir de um conjunto pré–determinado de valores como, por exemplo, 2, 5 ou 10 km.

Para figurar na lista de documentos recuperados em resposta a uma consulta, um documento deve atender a dois pré–requisitos: (1) ser compatível com o escopo geográfico da consulta; e (2) oferecer informações relacionadas aos produtos ou serviços descritos pelas palavras–chave informadas no objeto da consulta. Existem basicamente três formas de se lidar com esses requisitos, cuja origem encontra-se no método utilizado para adquirir a coleção de documentos onde as buscas são realizadas. Vamos ilustrá-las tomando como exemplo os serviços comerciais mais representativos de cada uma: *Google Local*[1] e *Yahoo! Local*[2], versões "geográficas" do *Google* e do *Yahoo!*, e *Geosearch*, uma máquina de busca especializada em busca local, que esteve disponível na Internet de abril de 2000 a março de 2002, sendo uma das pioneiras na área.

Os documentos utilizados pelo *Yahoo! Local* são registros provenientes de bancos de dados comerciais de cadastro de empresas, as tradicionais páginas amarelas, como a *US List*, comercializada pela *InfoUSA*[3], que contém cerca de 14 milhões de registros. Em contraste, o *Geosearch* utilizava apenas documentos coletados na *Web*, mais especificamente um sub–conjunto contendo localizadores geográficos. Já o *Google Local* utiliza uma abordagem mista, onde documentos da *Web* são correlacionados a registros provenientes de bancos de dados comerciais e páginas amarelas *on–line*. O elo entre essas fontes de dados é a presença de endereços ou números de telefone dos registros comerciais no texto dos documentos da *Web*.

Para atender ao primeiro requisito, isto é, determinar se um documento é compatível com o escopo geográfico da consulta, é preciso que ele tenha passado por uma etapa de pré–processamento, onde localizadores geográficos são encontrados (*geoparsing*) e então convertidos em coordenadas geográficas (geocodificação). Se as coordenadas determinarem um ponto dentro da área coberta pelo escopo da consulta, o documento é considerado compatível. No caso do *Google Local* e *Yahoo! Local*, o processo de *geopar-*

[1] `http://local.google.com`
[2] `http://local.yahoo.com`
[3] `http://www.infousa.com`

sing é trivial, pois os localizadores geográficos encontram-se em registros estruturados. Já no *Geosearch*, era preciso utilizar métodos mais apurados para detectá–los, uma vez que eles poderiam aparecer em qualquer posição dentro do texto de um documento da *Web* e em formatos diversos. Em todos eles, porém, é necessário contar com um banco de dados georreferenciados para que se possa atribuir coordenadas geográficas a um documento.

Em relação ao segundo requisito, todos os produtos utilizam técnicas de recuperação de informação envolvendo o casamento de palavras–chave para tentar recuperar documentos contendo informações relevantes ao usuário. No *Yahoo! Local*, o conjunto de palavras–chave que representa o objeto de interesse é pesquisado nos documentos, em campos como título e descrição, e também nas palavras–chave atribuídas às categorias nas quais foram *manualmente* classificados. No *Geosearch*, a pesquisa ocorria apenas no texto dos documentos, enquanto no *Google Local* tanto o texto dos documentos provenientes de registros estruturados quanto o texto dos documentos da *Web* associados a eles são pesquisados. Obviamente, assim como ocorre nas máquinas de busca, não há garantias de que, dado um conjunto de palavras–chave, a necessidade de informação do usuário seja completamente atendida.

Atendidos os pré–requisitos, uma lista de documentos ordenados por relevância é retornada como resposta à consulta. Uma gama variável de informações, tais como distância ao centro de busca, horário de funcionamento, preço médio, formas de pagamento, resenhas e classificações são em geral apresentadas, juntamente com o nome, descrição e dados de contato do local onde o serviço ou produto de interesse é comercializado. Com o auxílio de um mapa, por vezes contando até com imagens de satélite, o usuário pode visualizar a posição do local que procura, receber instruções de como chegar e mesmo encontrar locais na redondeza que comercializam outros produtos ou serviços (Figura 2.1).

2.2 Atribuição de Contexto Geográfico

Segundo Amitay et al. (2004), existem dois tipos de contexto geográfico que se pode atribuir a um documento da *Web*: o contexto da *origem* (*source*) e o contexto do *alvo* (*target*). O contexto geográfico da origem está relacionado à localização física dos servidores onde o documento está hospedado. Para tal, vale–se de evidências oriundas da infra–estrutura da Internet, como bases de registro de domínios, dados sobre a localização de roteadores e a estrutura hierárquica presente em nomes de domínio. O contexto geográfico do alvo é determinado pelo conteúdo do documento, a partir da identificação de elementos textuais considerados fontes de evidência de contexto geográfico —

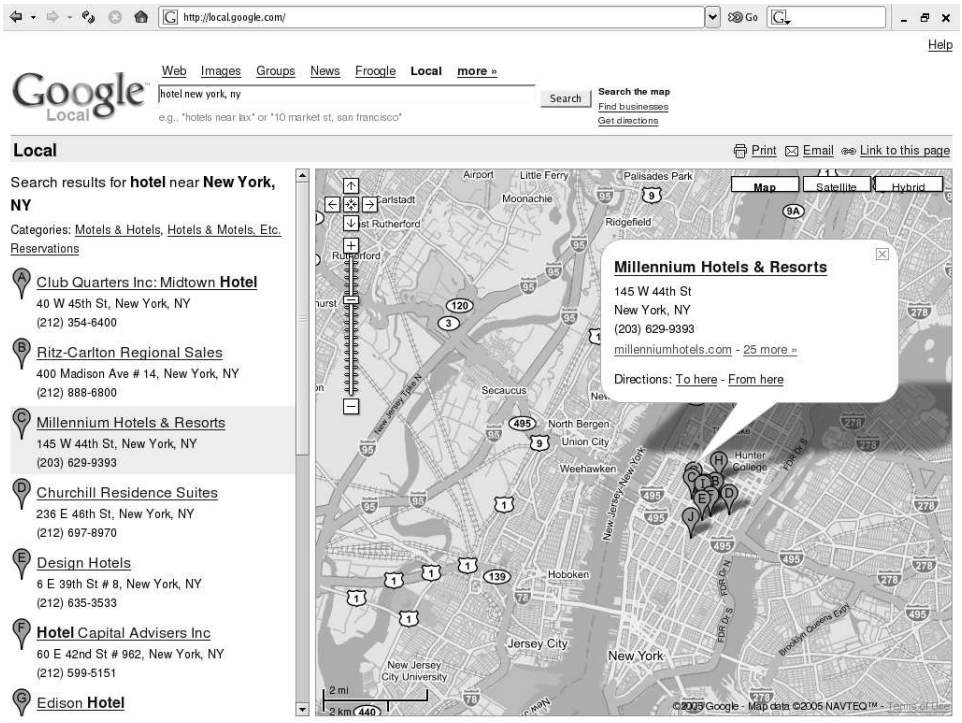

Figura 2.1: Visualização em mapa do resultado da consulta 'hotel new york, ny' no *Google Local*.

endereços, códigos postais e números de telefone, nomes de lugares, *hyperlinks* e metadados. O contexto do alvo está relacionado ao assunto tratado no documento. Em ambos os casos, entende-se por atribuição de contexto a designação de um conjunto representativo de coordenadas geográficas a um documento, obtidas mediante processo de geocodificação das fontes de evidência. McCurley (2001) faz um apanhado geral sobre as fontes de evidência utilizadas para atribuir os dois tipos de contexto geográfico a documentos da *Web*.

Buyukkokten et al. (1999) utilizam em seu trabalho a atribuição de contexto geográfico da origem a *sites* da *Web*. A partir da geocodificação de números de telefone e códigos postais (*zipcodes*) obtidos com consultas 'whois' a bases de registro de domínios, é criado um relacionamento entre endereços IP e as coordenadas médias de latitude e longitude que representam a localização geográfica dos *sites*. Inferências sobre a abrangência geográfica do *site*, por eles denominada escopo geográfico, são feitas combinando–se essa informação aos *hyperlinks* direcionados a documentos do *site*, originados de documentos hospedados em locais diversos. O jornal *New York Times*, por exemplo, tem um escopo mais abrangente do que o jornal *San Francisco Chronicle*, pois é apontado por documentos de *sites* espalhados por todo o território americano. O artigo apresenta ainda uma ferramenta de visualização que exibe, em um mapa, as localidades alcançadas pelo escopo geográfico do *site* selecionado.

Embora com a estratégia proposta por Buyukkokten et al. seja possível analisar o escopo geográfico de um *site* da *Web*, ela não é apropriada para recuperar um documento com base em critérios geográficos, uma vez que é incorreto inferir que seu conteúdo está relacionado ao lugar onde está hospedado. Para sanar essa deficiência, Ding et al. (2000) estendem esse trabalho, introduzindo técnicas para reconhecimento automático do contexto geográfico de *documentos* da *Web*. Além de considerar o escopo geográfico dos *sites*, determinado pelos *hyperlinks*, a distribuição geográfica dos nomes de lugares mencionados no texto de um documento também é observada, levando a uma abordagem mista de atribuição de contexto geográfico, que utiliza tanto elementos da origem quanto do alvo. O protótipo de uma máquina de busca foi construído a partir da indexação dos documentos coletados de 436 jornais *on–line*. Especificando–se como consulta um código postal juntamente com um conjunto de palavras–chave, os documentos mais relevantes, compatíveis em escopo geográfico e conteúdo são retornados.

McCurley (2001) investiga uma variedade de fontes de evidência a partir das quais se pode atribuir contexto geográfico a um documento. Ele aborda o problema de indexação e, principalmente, navegação por critério geográfico, apresentando um sistema experimental onde a função '*where*' pode ser executada pelo usuário a qualquer momento, durante a utilização de um navegador *Web*, para visualizar em um mapa os lugares associados ao documento corrente. No mapa há ainda pontos em destaque que correspondem a lugares associados a outros documentos que, pelo contexto geográfico, estão próximos ao documento que está sendo visualizado. Clicando em qualquer um dos pontos destacados no mapa, uma lista de destinos é apresentada ao usuário. Ao escolher um item da lista, o navegador é atualizado com o documento selecionado, e o mapa passa a exibir os lugares identificados no novo documento.

Até o presente momento, a existência de trabalhos que descrevem a utilização de expressões de posicionamento como forma de se atribuir contexto geográfico a documentos *Web* não é de nosso conhecimento. Entretanto, existem trabalhos envolvendo o uso das relações espaciais na especificação de consultas e, principalmente, trabalhos sobre a atribuição de contexto geográfico a documentos da *Web* pela identificação de nomes de lugares, e que são relevantes para o assunto aqui tratado.

Em (Silva et al., 2004), (Heinzle et al., 2003) e (Sanderson e Kohler, 2004) é destacada a necessidade das máquinas de busca entenderem e fazerem uso das relações espaciais como forma de reconhecer as consultas geográficas e para melhorar a qualidade das respostas a essas consultas. Já Rodríguez-Tastets (2002) apresenta uma abordagem baseada no conhecimento geográfico (*knowledge–based approach*) para aumentar a capacidade de resposta das máquinas de busca atuais em consultas expressas por meio de operadores espaciais tais como 'em', 'norte', 'oeste', 'perto'. Como exem-

plo, é citada a consulta 'encontre hotéis em Pucón, Chile, e nas cidades adjacentes'. Para atender a esse tipo de consulta é descrito um modelo para organização e derivação das relações espaciais baseadas na estrutura hierárquica do espaço, organizado em regiões, e considerando as inter–relações das regiões conectadas. A idéia geral é mapear o nome de regiões e lugares dentro dessa estrutura semântica de tal modo que os documentos da *Web* associados a esses lugares não necessitem incorporar propriedades geométricas em seu texto, uma vez que as análises envolvendo os relacionamentos topológicos, cardinais e de distância são favorecidos ao se utilizar o raciocínio espacial. O trabalho apresenta diretrizes de como esse modelo pode ser usado para estender as técnicas atuais de busca a documentos da *Web*, sem no entanto abordar como os nomes de lugares podem ser mapeados para a estrutura semântica do espaço proposto.

Dentre as fontes de contexto geográfico, as mais desafiadoras são sem dúvida os nomes de lugares. Endereços, códigos postais e números de telefone são geralmente regidos por regras, padrões e convenções, o que os torna facilmente identificáveis em meio ao texto de um documento utilizando–se, por exemplo, regras sintáticas expressas por meio de expressões regulares. Devido à sua imensa diversidade, os nomes de lugares, ao contrário, não podem ser identificados dessa forma. Além disso, existem vários lugares distintos com o mesmo nome, o que gera a necessidade do emprego de métodos para determinar o exato lugar ao qual um nome se refere, os chamados métodos de resolução de ambigüidade. Repositórios de nomes de lugares, chamados *gazetteers* (Hill, 2000), são invariavelmente empregados nas tarefas de resolução de ambigüidades e identificação dos nomes de lugares, esta muitas vezes realizada em conjunto com *taggers* genéricos como o GATE (Cunningham et al., 2002), *softwares* capazes de identificar nomes de entidades como pessoas, organizações e locais em textos em linguagem natural. Amitay et al. (2004) classificam as ambigüidades como *geo/geo* (Paris, França *versus* Paris, Texas) ou *geo/não–geo* (Tiradentes, Minas Gerais *versus* Tiradentes, o inconfidente).

Os trabalhos de Amitay et al. (2004) e Zong et al. (2005) preocupam–se em (1) identificar nomes de lugares; (2) resolver eventuais ambigüidades; e (3) designar um contexto geográfico aos documentos a partir dos nomes de lugares nele mencionados. Esses trabalhos apresentam várias técnicas para retirada de ambigüidade, geralmente baseadas em pistas textuais encontradas próximas aos nomes e na correlação com outros nomes não–ambígüos ou com a ambigüidade resolvida, no mesmo documento. As estratégias empregadas por Amitay et al. atingiram 81,7% de precisão, enquanto as de Zong et al. chegaram a 88,9%. Com relação à atribuição de contexto geográficos, Amitay et al. utilizam no máximo quatro nomes de lugares para representar a abrangência do documento como um todo, o que denominaram *"foco" geográfico*. Já Zong et al. preocupam–se em atribuir, para segmentos de texto em um documento, um

lugar representativo dentre aqueles mencionados. Por exemplo, se em um segmento são mencionados as cidade de Belo Horizonte, Contagem, Ribeirão das Neves e Santa Luzia, a Região Metropolitana de Belo Horizonte pode ser um nome apropriado para descrever o contexto geográfico desse segmento.

Como veremos na Seção 3.3, neste trabalho foi implementado um programa para identificar expressões de posicionamento em documentos textuais, onde há também a necessidade de se identificar nomes, porém não de lugares relacionados às divisões geo–políticas como é de praxe, e sim nomes de locais intra–urbanos, como pontos turísticos e edificações de destaque. Em nosso programa, o uso de um *gazetteer* é dispensável, e o emprego de relações espaciais para contextualizar as ocorrências de nomes de locais, de forma semelhante à utilizada em (Woodruff e Plaunt, 1994; Rauch et al., 2003), praticamente elimina as ambigüidades do tipo *geo/não–geo*.

Capítulo 3

Expressões de Posicionamento

Os seres humanos estão em geral acostumados a lidar com imprecisão e ambigüidade quando o assunto é localização. Em linguagem natural, habitualmente nos referimos a lugares e locais por intermédio de um nome. Se precisamos mencionar, seja em um texto ou durante uma conversa, algum local pouco conhecido ou cujo nome possa ser confundido com o de outro local, recorremos a aproximações e posicionamentos relativos, para tentar nos fazermos entender. Dessa forma, geralmente indicamos a localização de nosso interesse mediante relacionamento com outro local, chamado *ponto de referência*, confiando que esta combinação possa transmitir uma descrição aproximada, não obstante útil e compreensível, da localização exata desejada.

Como conseqüência, textos em linguagem natural, incluindo os documentos da *Web*, freqüentemente contêm esclarecimentos e indicações sob a forma de alusões a pontos de referência que buscam situar no espaço um local de interesse. Essas alusões estão embutidas em estruturas textuais aqui denominadas *expressões de posicionamento*, formadas por relações espaciais e pontos de referência. Neste capítulo, através de uma série de experimentos baseados em um estudo de caso da cidade de Belo Horizonte, procuramos caracterizar essa fonte de evidência de contexto geográfico e compreender melhor as informações representadas por ela. Antes, porém, definimos na Seção 3.1 alguns conceitos básicos e a terminologia utilizada ao longo do texto.

3.1 Conceitos e Terminologia

Ponto de Referência Lugar ou local facilmente reconhecível, em geral amplamente conhecido, que as pessoas utilizam para julgar onde estão. Cabe aqui uma distinção entre os significados adotados para lugar e local. Tomamos por *lugar* as divisões geopolíticas que formam uma hierarquia de relacionamentos do tipo *todo–parte*, quais sejam, bairros, cidades, micro–regiões, macro–regiões, estados,

países e assim por diante. Já por *local* entenda–se uma estrutura visual cujos limites podem ser facilmente identificados. Um local está contido em um lugar.

Pela definição, um lugar pode ser utilizado como um ponto de referência, fato ilustrado pelos seguintes exemplos: "Estou morando na **Savassi**"; "o município de Betim está situado próximo a **Belo Horizonte**". Entretanto, de particular interesse para este trabalho são os locais empregados como ponto de referência, especialmente os locais *intra–urbanos*.

Relação Espacial Expressão em linguagem natural utilizada para descrever relacionamentos entre entidades geográficas. Apesar de serem amplamente estudadas no contexto dos SIG (veja o levantamento feito por Papadias e Sellis (1994)), onde são definidas como funções ou operadores entre objetos espaciais, pouco se sabe a respeito das relações espaciais em linguagem natural, como as encontradas em documentos da *Web*. Não obstante, utilizamos uma taxonomia oriunda dos trabalhos de Pullar e Egenhofer (1988), Egenhofer e Franzosa (1991) e Guting (1994) para agrupá–las. Uma relação espacial pode ser classificada como topológica, métrica, direcional ou *fuzzy*.

As relações *topológicas* representam o grau de conectividade entre entidades geográficas. São relações que descrevem os conceitos de vizinhança, incidência e sobreposição, mantendo–se invariantes ante transformações como escala e rotação (Guting, 1994). Tradicionalmente, as relações topológicas têm sido encaradas como as estruturas espaciais mais abstratas, consideradas essencialmente qualitativas. De fato, essas relações são sem dúvida as mais estudadas dentro do contexto dos SIGs. Em linguagem natural, expressões como 'dentro de', 'em', 'no primeiro andar de', 'embaixo de' e 'no coração de' foram classificadas como topológicas.

As relações *métricas* descrevem proximidade de forma quantitativa e possuem a seguinte forma geral: 'a X Y de'. O grau de proximidade entre as entidades é dado pelos parâmetros de valor (X) e de grandeza (Y). Exemplos: 'a 10 km de', 'a oitocentos metros de', 'a vinte minutos de' e 'a 3 quadras de'.

As relações *direcionais* descrevem orientação e ordem. A orientação é determinada por direções como as cardinais ('ao norte de', 'a leste de') enquanto que expressões como 'em frente a', 'atrás de' e 'ao lado de' indicam a ordem total ou parcial entre os objetos espaciais (Freeman, 1975).

Por fim, as relações *fuzzy* descrevem proximidade mediante o emprego de termos essencialmente qualitativos, dependentes do contexto e, portanto, imprecisos, como 'perto de', 'na vizinhança de' e 'próximo a' (Pullar e Egenhofer, 1988).

Expressão de Posicionamento Construção em linguagem natural utilizada tanto na fala quanto na escrita para expressar, por meio de um par ⟨*relação espacial, ponto de referência*⟩, a posição relativa no espaço de um *objeto de interesse* (veja definição a seguir).

A efetividade e precisão com que uma expressão de posicionamento é capaz de descrever uma localização depende de vários fatores, os mais notórios sendo:

- o nível de conhecimento que o interlocutor possui a respeito da localização do ponto de referência — muitas pessoas sabem que o *Empire State Building* fica em Nova York, mas provavelmente apenas alguns nova–iorquinos saberão que ele se localiza na *Fifth Avenue*, número 350, entre as ruas 33 e 34, a um quarteirão da *Penn Station*.
- emprego de sinônimos — um mesmo local pode ser referenciado por vários nomes. Uma pessoa pode compreender o significado de estar 'perto do Mineirão' e desconhecer o de estar 'perto do Estádio Governador Magalhães Pinto', quando na verdade essas expressões de posicionamento são equivalentes. Uma busca por um local em uma aplicação de RIG deveria recuperar documentos que mencionam quaisquer de seus nomes.
- conflito de nomes — o nome de um ponto de referência pode ser utilizado, em função do contexto, para designar diferentes locais ou mesmo possuir outros significados que não o geográfico, o que gera ambigüidades do tipo *geo/geo* e *geo/não–geo*. Além disso, palavras genéricas, como 'aeroporto' e 'centro', servem para designar locais em vários lugares do Brasil e do mundo.
- contexto de utilização das relações espaciais — o sentido geográfico de uma relação espacial varia de acordo com as características do ponto de referência, como tamanho, forma e localização. A relação 'perto de' nas expressões 'perto de Londres' e 'perto do Big Ben' possui dimensões completamente distintas; 'a 25 minutos de carro do centro' pode ser bem diferente em termos de distância caso estejamos nos referindo ao centro de uma cidade grande ou de um vilarejo.
- regionalismo e subjetividade — os termos empregados em uma relação espacial variam de lugar para lugar, com algumas expressões sendo mais comuns que outras, inclusive em um mesmo país ou região. Além disso, a descrição de um relacionamento espacial como uma relação espacial em linguagem natural é uma tarefa subjetiva. Alguns relacionamentos englobam outros e vários termos semanticamente equivalentes podem ser empregados.

Objeto de Interesse Entidade ou evento associado a uma localização descrita por uma expressão de posicionamento. Como exemplos temos: 'o parque localiza–se próximo ao centro da cidade' (entidade) e 'as explosões ocorreram a menos de um quilômetro do Parlamento' (evento). São de especial importância para este trabalho os objetos de interesse do tipo entidade que são locais, os quais denominamos *locais de interesse*.

Além de explicitamente, como nos exemplos anteriores, um objeto de interesse pode aparecer em um documento da *Web* de forma implícita, como no *site* de um hotel cuja localização é descrita por uma expressão de posicionamento sem que haja menção a seu nome.

Localizador Geográfico Conjunto de símbolos e nomes, geralmente regidos por regras, padrões e convenções, associados a coordenadas geográficas que determinam um ponto, segmento de reta ou área na superfície terrestre. Pode ser do tipo *direto* ou *indireto*.

Os localizadores geográficos indiretos funcionam como rótulos, uma forma mais conveniente aos seres humanos de expressar posições no espaço. Endereços, códigos postais, números de telefone, nomes de lugar e expressões de posicionamento são exemplos de localizadores geográficos indiretos. Já os localizadores diretos são uma representação imediata de coordenadas geográficas, como os pontos de latitude e longitude `S 20º07.403' W 40º57.954'`.

Em aplicações de RIG, os localizadores geográficos presentes em documentos da *Web* são ditos serem *fontes de evidência de contexto geográfico*, *fontes de contexto geográfico* ou simplesmente *evidências geográficas*.

Referência Geográfica Par ⟨*objeto de interesse, localizador geográfico*⟩. O subconjunto de referências geográficas representado pelo par ⟨*local de interesse, expressão de posicionamento*⟩ é enfatizado neste trabalho. Exemplos: ⟨'Mineirão', 'perto da Lagoa da Pampulha'⟩, ⟨'Hotel Ibis', 'em frente ao Aeroporto de Congonhas'⟩.

3.2 Estudo de Caso: Belo Horizonte

Nosso conhecimento prévio acerca das expressões de posicionamento presentes em documentos da *Web* era limitado e pouca informação pôde ser obtida na bibliografia. Dessa forma, decidimos realizar um estudo de caso com o objetivo de coletar dados e produzir informações relevantes a respeito dessas expressões. Como descrito nas próximas seções, os experimentos realizados concentram–se na identificação, classificação

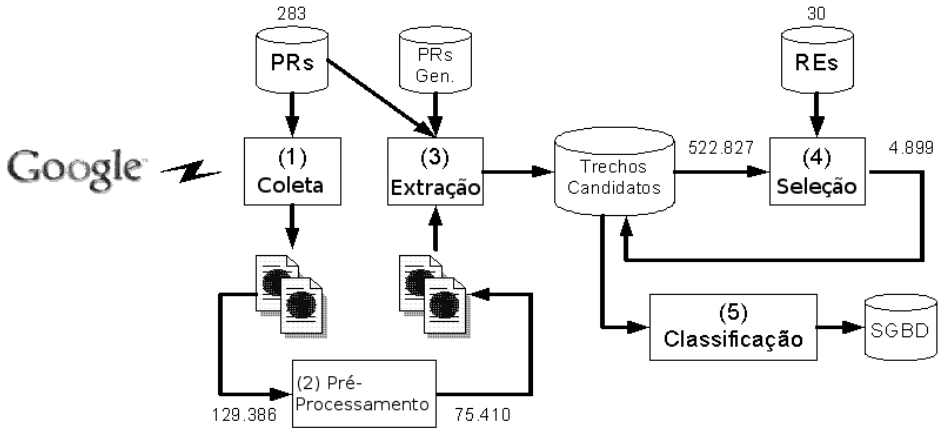

Figura 3.1: Visão geral do processo de aquisição e classificação de expressões de posicionamento do estudo de caso.

e análise de expressões de posicionamento utilizadas em documentos da *Web* como localizadores geográficos de objetos de interesse relacionados à cidade de Belo Horizonte. BH, como é carinhosamente chamada pelos belo–horizontinos, é a capital do estado de Minas Gerais e o município sede da terceira maior região metropolitana do Brasil, que abriga quase 4.9 milhões de habitantes[1].

Belo Horizonte foi a cidade escolhida devido à ampla variedade de dados disponíveis, obtidos do seu SIG (Borges e Sahay, 2000). O SIG de Belo Horizonte utiliza um completo e preciso banco de dados de endereçamento urbano, que mantém mais de 300 classes de objetos, incluindo um conjunto de 420.000 endereços individuais, georreferenciados como pontos. Naturalmente, os experimentos foram realizados em português, mas a mesma abordagem poderia ter sido empregada em outras línguas, como inglês e espanhol.

3.2.1 Identificação de Expressões de Posicionamento

Para alcançar nosso objetivo, foi criada uma pequena lista com 283 nomes, oficiais e alternativos, de pontos de referência representativos de Belo Horizonte. A maioria dos pontos de referência designados localizam–se na região da Pampulha, local turístico, propenso ao lazer e, sobretudo, na região Centro–Sul, principal centro administrativo, financeiro, comercial e cultural da cidade. A lista inclui locais como prédios públicos, atrações turísticas, edificações de destaque, hospitais, estações de transporte coletivo, faculdades, igrejas, obras de arte, praças, entre outros.

[1] `ftp://ftp.ibge.gov.br/Estimativas_Projecoes_Populacao/Estimativas_2005/UF_Municipio.zip`

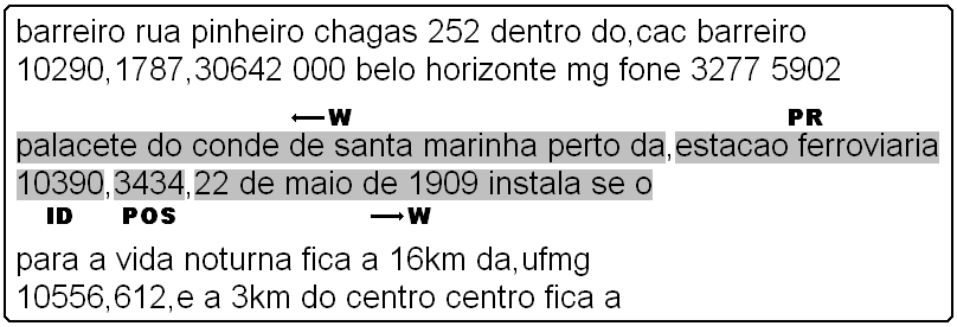

Figura 3.2: Exemplo da saída do extrator.

Utilizando os serviços da API *Web* do *Google*[2], os 283 pontos de referência da lista foram submetidos de forma automática a essa máquina de busca, sob a forma de consultas por frase, como ilustra o passo 1 da Figura 3.1. Todos os documentos retornados como resposta foram coletados. Após a remoção de documentos em outros formatos que não fossem HTML, tais como *Portable Document Format* (PDF), *PostScript* (PS), MS–Word e MS–Powerpoint, e de documentos repetidos, identificados com base na URL, obtivemos uma coleção de teste com 75.410 documentos. Essa base textual foi ainda pré–processada para substituição das *tags* HTML pelo marcador '<>', remoção dos acentos ortográficos, substituição de caracteres de controle tais como `TAB`, `FF` e `CR` por espaço e eliminação de espaços consecutivos (passo 2 da Figura 3.1).

Com a coleção de documentos pronta para utilização, o próximo passo foi implementar um programa para identificar padrões textuais. Esse programa, chamado de *extrator de padrões textuais*, ou simplesmente *extrator*, é baseado no algoritmo de casamento eficiente de cadeias de caracteres proposto por Aho e Corasick (1975). Esse algoritmo e suas variações são utilizados atualmente em diversos *softwares*, incluindo a ferramenta *grep*, presente na maioria das variantes do sistema operacional UNIX, como Linux, BSD e Solaris.

O extrator recebe como entrada uma coleção de documentos em formato texto e um arquivo contendo os padrões textuais que se deseja encontrar. Os nomes dos 283 pontos de referência enviados à máquina de busca, quando da coleta dos documentos, acrescidos de nomes de locais genéricos como 'centro', 'aeroporto', 'zoológico' e 'rodoviária', foram utilizados como padrões. A pesquisa é feita em tempo diretamente proporcional ao tamanho dos textos dos documentos, independente do número de padrões (passo 3 da Figura 3.1).

Como ilustrado na Figura 3.2, para cada casamento bem sucedido o extrator imprime o ponto de referência encontrado precedido por W palavras à sua esquerda, formando o que denominamos *trechos candidatos* — passagens textuais candidatas a

[2] `http://www.google.com/apis/`

conter uma expressão de posicionamento. Junto a cada trecho candidato, o extrator imprime ainda o identificador do documento de onde o trecho foi extraído, a posição no texto onde ele se encontra e W palavras à direita do ponto de referência para melhor contextualizá–lo. Nesse experimento utilizamos $W = 8$, sendo que 522.827 trechos foram gerados.

Selecionamos palavras *à esquerda* do ponto de referência para formar os trechos candidatos pois é nessa parte do texto que se espera que ocorra a relação espacial que o acompanha, formando uma expressão de posicionamento. Isso segundo a construção sintática mais comum, na língua portuguesa, para esse tipo de expressão. Porém, nada impede que inversões sejam empregadas, posicionando–se a relação espacial após o ponto de referência, conforme o seguinte exemplo, retirado de um dos documentos da coleção: *"Contam os baianos que, na cidade, existem 365 igrejas, uma para cada dia do ano. Como, no Terreiro de Jesus, a imponente* ***Catedral Basílica***. ***Próximo a*** *ela, está a Igreja de São Francisco de Assis, com arabescos em madeira banhada a ouro."*

Na passagem acima, se 'Catedral Basílica' for um ponto de referência, uma busca pela relação espacial 'próximo de' no trecho candidato correspondente não será bem sucedida. Já na forma direta — "A Igreja de São Francisco de Assis fica **próxima à** imponente **Catedral Basílica**, no Terreiro de Jesus" — a expressão de posicionamento ⟨'próximo de', 'Catedral Basílica'⟩ será encontrada normalmente.

O próximo passo rumo à identificação de expressões de posicionamento foi selecionar, dentre os trechos candidatos gerados pelo extrator, aqueles que continham uma ou mais relações espaciais. Para identificar as relações espaciais, definimos um conjunto de expressões regulares estendidas (IEEE, 2001) capazes de representar 30 relações espaciais no *formato básico*, além de dezenas de variações. Tais variações incluem gênero (feminino/masculino), número (singular/plural), advérbios (ex.: 'muito'), abreviaturas (ex.: 'aprox.'), regionalismos, entre outros. As expressões regulares definidas encontram–se listadas no Apêndice A. Como exemplo, temos abaixo a expressão regular para a relação espacial 'km de':

```
'\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| aprox(
?:imadamente|.?)| uns| cerca de| (?:pouco )?m(?:ais|enos) de| quase| exatos|
 mais ou menos| [+][-])? (?:[[:alpha:]]+ |[[:digit:]][[:digit:],.]* ?)km d(?
e distancia d[oae]|[oae])) }'
```

Essa expressão é capaz de reconhecer relações tão diversas quanto 'a vinte km do', 'a exatos 4,7 km da', 'a uma distancia de aproximadamente cinco km de' e 'dista cerca de 100 km de'.

As relações espaciais foram escolhidas dentre as mais representativas de cada uma das categorias descritas na Seção 3.1, junto à definição de 'Relação Espacial'. É importante, porém, ressaltar que a formação de um conjunto completo, definitivo, não

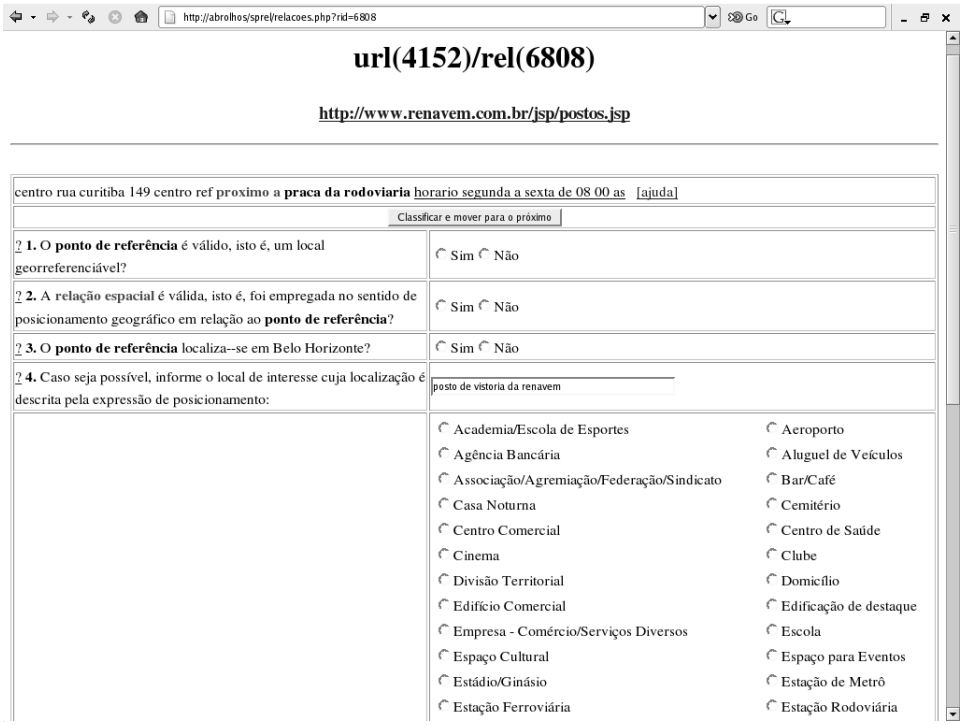

Figura 3.3: Interface *Web* para classificação das expressões de posicionamento.

era uma preocupação. A relação espacial 'em' e suas variações 'no' e 'na' foram propositalmente excluídas pois, em experimentos preparatórios, percebemos que elas são extremamente comuns e estão em grande parte associadas a divisões territoriais como bairros e cidades, enquanto nossa ênfase são os pontos de referência intra–urbanos. Além disso, percebemos que elas aparecem com grande freqüência em contextos diversos, que não aqueles descrevendo a localização de um local de interesse.

Com o auxílio da ferramenta *grep*, selecionamos 4.889 trechos candidatos contendo pelo menos uma relação espacial. A ferramenta foi escolhida pelo fato de as relações espaciais estarem descritas sob a forma de expressões regulares e pela sua excelente performance (passo 4 da Figura 3.1).

3.2.2 Classificação das Expressões de Posicionamento

Os 4.889 trechos candidatos selecionados tiveram seus dados carregados em um banco de dados e foram manualmente inspecionados, um a um, por cinco pessoas do nosso grupo de pesquisa (passo 5 da Figura 3.1). A interface *Web* exibida na Figura 3.3 foi especialmente desenvolvida para essa atividade. O processo de classificação consistiu das seguintes etapas:

- Verificar se a expressão identificada como um ponto de referência está correta; ajustar no banco de dados se necessário. O ajuste pode ocorrer quando o nome de

um ponto de referência está contido no nome de outro que não esteja presente na lista de pontos de referência utilizada. Um outro caso comum é quando pontos de referência aparecem próximos, e apenas um deles consta na lista. Veja os exemplos abaixo, nos quais o segundo trecho encontra–se ajustado. Procedendo dessa forma, acabamos por identificar novos pontos de referência.

- De "Parque Municipal, **defronte à** Escola de Medicina da **UFMG**" para "Parque Municipal, **defronte à Escola de Medicina da UFMG**";
- De "Avenida Raja Gabaglia, **próximo à** Polícia Federal, sentido **BH Shopping**" para "Avenida Raja Gabaglia, **próximo à Polícia Federal**, sentido BH Shopping";
- De "Av. Paraná, **em cima do** Banco do Brasil - **Centro**" para "Av. Paraná, **em cima do Banco do Brasil** - Centro".

- Preencher um formulário com informações a respeito da expressões de posicionamento. O formulário é composto por sete itens, com uma ordem específica de preenchimento, e um espaço adicional para que se possa fazer qualquer tipo de observação com relação à classificação do trecho. Os itens são listados a seguir, comentados:

  1. "*O ponto de referência é válido, isto é, um local georreferenciável? R: Sim/Não. Em caso positivo, prossiga para o próximo item*". Algumas expressões, selecionadas como pontos de referência, são na verdade homônimos: possuem a mesma pronúncia e grafia, com significado diferente. Exemplos: "O carioca passa num carrão **perto de** um **mineirinho**, em cima de uma carroça"; "**depois da** revolta do Terceiro Regimento de Infantaria, **centro** do movimento conhecido como Intentona Comunista".
  2. "*A relação espacial é válida, isto é, foi empregada no sentido de posicionamento geográfico em relação ao ponto de referência? R: Sim/Não. Em caso positivo, prossiga para o próximo item*". Algumas expressões, selecionadas como relações espaciais, são empregadas em contextos diferentes daquele esperado. Exemplos: "**Depois do** trabalho fui ao **BH Shopping**"; "Primus, cada vez mais **perto de** você! **Aeroporto Internacional de Confins** Tel.: (31) 3689-2044"; "de um esquema de corrupção **dentro da prefeitura** que o PT sabia e não queria expor".
  3. "*O ponto de referência localiza–se em Belo Horizonte? R: Sim/Não. Em caso positivo, prossiga para o próximo item*". O objetivo é selecionar locais em Belo Horizonte.

4. "*Caso seja possível, informe o local de interesse cuja localização é descrita pela expressão de posicionamento. R: Livre. Prossiga para o próximo item se o local de interesse for informado*".
5. "*Marque o tipo de local que melhor representa o local de interesse. R: um dentre os tipos disponíveis. Prossiga para o próximo item*". É interessante saber quais os tipos mais representativos de locais de interesse cuja localização é descrita por uma expressão de posicionamento.
6. "*Caso seja possível, informe o endereço do local de interesse. R: Livre. Prossiga para o próximo item se o endereço for informado*". Essa informação será utilizada na Seção 3.4, para geocodificar o local de interesse.
7. "*Onde o endereço foi encontrado? R: Na própria página/Outros Locais*". Com essa informação, podemos descobrir quantos objetos de interesse possuem a sua localização descrita apenas por uma expressão de posicionamento.

Após a execução dos procedimentos descritos, os trechos identificados como sendo referentes a Belo Horizonte tiveram seus pontos de referência classificados nos mesmos tipos disponíveis para os locais de interesse. Além disso, o endereço desses pontos, quando disponíveis, também foram registrados. É importante notar que um mesmo local pode aparecer tanto como um ponto de referência quanto como um local de interesse. Dessa forma, parte do esforço utilizado durante a classificação e pesquisa do endereço para os locais de interesse foi reaproveitado.

3.2.3 Análise das Expressões de Posicionamento

Como poderá ser percebido pelas figuras e tabelas a seguir, muitas informações a respeito das expressões de posicionamento foram obtidas a partir da interpretação dos dados gerados pela etapa de classificação. Os resultados mais importantes foram incluídos nesta seção, alguns de forma resumida, sendo que informações mais detalhadas podem ser encontradas no Apêndice B.

A Tabela 3.1 exibe um resumo da classificação realizada nos trechos candidatos. *Trechos inválidos* são aqueles em cujo questionário há uma resposta 'Não' para o item 1 *ou* 2, isto é, não contêm uma expressão de posicionamento. Já os *trechos válidos* contêm uma expressão de posicionamento, ou seja, 'Sim' foi a resposta aos itens 1 *e* 2 do questionário. Portanto, como resultado da classificação, 89,43% ou 4.372 trechos foram considerados válidos enquanto que 517 trechos (10,57%), foram considerados inválidos. A ocorrência de trechos inválidos deve-se, em grande parte, ao fato de

Tabela 3.1: Resumo da classificação dos trechos candidatos. As porcentagens referem–se ao valor do item no nível superior.

Caso	Ocorrências		Trecho típico
Total	4889	100,00%	
Trecho inválido	517	10,57%	"...um projeto reestruturador dentro da UFMG."
Trecho válido	4372	89,43%	
em outros municípios	3463	79,21%	"...perto da prefeitura..."
em Belo Horizonte	909	20,79%	
cujo objeto de interesse *não* é um local	165	18,15%	"...o carro foi encontrado nas imediações da Praça Raul Soares."
cujo objeto de interesse *é* um local	744	81,85%	"...o hotel localiza–se a duas quadras do Minas-centro."
sem endereço	189	25,40%	
com endereço	555	74,60%	
encontrado no documento	364	65,59%	
encontrado de outra forma	191	34,41%	

que algumas relações espaciais como 'antes de', 'dentro de' e 'depois de' são também empregadas em contextos diferentes do geográfico, com outra interpretação semântica.

Apesar de a coleta ter sido feita objetivando a recuperação de documentos contendo pontos de referência de Belo Horizonte, a quantidade de trechos de outros municípios foi quase quatro vezes maior. Em grande parte esse fenômeno deve–se ao emprego de termos genéricos durante a seleção de trechos candidatos. A palavra 'centro', por exemplo, além de existir como um local, geralmente o centro comercial de uma cidade, ainda compõe o nome de vários pontos de referência, como 'Centro de Convenções Anhembi', que fica em São Paulo.

Locais de interesse não são as únicas entidades cuja localização é descrita em documentos da *Web* por meio das expressões de posicionamento. Em 18,15% dos trechos relacionados a Belo Horizonte, o objeto de interesse não é um local e sim algo circunstancial como um evento (ex.: um arrombamento ou manifestação) ou uma narração de fatos (ex.: 'estávamos reunidos, aguardando o jogo em frente ao Mineirão'), essa última forma muito comum em *blogs*.

Apesar de nosso esforço em tentar identificar o endereço de todos os locais de interesse, para 25,40% dos locais, isso não foi possível. Nesses casos, o endereço não estava disponível no texto do próprio documento e havia muito pouca informação para que ele pudesse ser encontrado por outros meios, como buscas na *Web* ou consultas

a catálogo de endereços. Tipos de local para os quais isso acontece com freqüência incluem *imóveis* e *placas de publicidade*[3] pois, em geral, o endereço exato não existe ou não é fornecido para esses locais. Em seu lugar, há apenas indicações sobre a localização dadas por expressões de posicionamento, como nos exemplos a seguir:

- *(outdoor)* Av Carlos Luz – **Ao lado do Motel Sunny** – Sentido centro (Frontal)
- *(outdoor)* Avenida Pedro I – **em frente à Vila Olímpica** – sentido Centro (frontal)
- *(imóvel)* APTO – Vendo, São Lucas, 3 qts. c/ armários, sala p/ 2 amb, coz. c/ arm, DCE, garagem. **Próximo à Santa Casa**. 9952-7191/3466-7894. (noite)
- *(imóvel)* CASA – Vendo, Carlos Prates, 3 qts, terraço, barracão independente, garagem e loja. **Próximo ao Centro de BH**. Só R$ 60 mil. 9691-3190.

A Figura 3.4 exibe a distribuição, por categoria, das relações espaciais encontradas nos conjuntos de trechos válidos e inválidos, enquanto a Tabela 3.2 exibe as contribuições individuais das dez relações espaciais mais freqüentes. Com relação aos trechos válidos, não há uma diferença significativa entre as participações das relações métricas, direcionais e *fuzzy*, sendo que a última leva alguma vantagem sobre as outras duas. O destaque fica para a pequena participação das relações topológicas, com apenas 8,6%. Já nos trechos inválidos, as relações *fuzzy* e topológicas são dominantes, com uma inexpressiva marca de 0,19% registrada para as relações métricas e uma contribuição de 7,54% das relações direcionais. Esses dados nos levam a acreditar que as relações direcionais e, principalmente, as métricas são predominantemente utilizadas no contexto de uma expressão de posicionamento, enquanto que as relações *fuzzy* e topológicas são empregadas em contextos diversos.

Dentre os trechos válidos, existem os que foram classificados como contendo expressões de posicionamento em Belo Horizonte e aqueles onde as expressões de posicionamento referem–se a outros municípios. Uma comparação entre a distribuição das categorias de relações espaciais entre esses trechos pode ser visualizada na Figura 3.5. Diferente do que poderíamos supor, em Belo Horizonte a distribuição das relações espaciais não segue uma proporção semelhante àquela observada para os trechos válidos em geral — em Belo Horizonte, as relações topológicas, *fuzzy* e direcionais são mais utilizadas, em detrimento das relações métricas. O emprego de poucas relações como 'a ? (minutos|km|quilômetros)' pode ter ocorrido em função de Belo Horizonte, com uma área de 330,954 km^2, ser um município de tamanho bastante modesto, já que,

[3]Por placa de publicidade entenda-se os pontos de veiculação de publicidade externa com localização fixa, tais como, *outdoors*, *front lights* e fachadas de prédios.

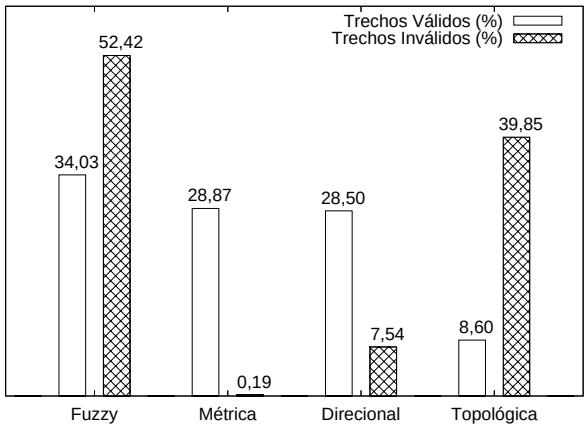

Figura 3.4: Distribuição das relações espaciais por categoria.

Tabela 3.2: Principais relações espaciais.

Trechos Válidos				Trechos Inválidos			
Cat.	Relação Espacial	Ocorrências		Cat.	Relação Espacial	Ocorrências	
F	próximo a	978	22.4%	T	dentro de	183	35.40%
D	em frente a	736	16.8%	F	depois de	114	22.50%
M	a ? km de	648	14.8%	F	antes de	90	17.41%
D	ao lado de	382	8.7%	F	perto de	31	6.00%
F	perto de	336	7.7%	D	ao lado de	17	3.29%
M	a ? minutos de	229	5.2%	D	atrás de	15	2.90%
M	a ? quilômetros de	216	4.9%	F	próximo a	14	2.71%
T	dentro de	216	4.9%	F	acima de	14	2.71%
D	atrás de	96	2.2%	T	em cima de	13	2.51%
F	nas proximidades de	90	2.1%	D	em frente a	7	1.35%
	SUB–TOTAL	3927	89.8%		SUB–TOTAL	498	96.32%
	TOTAL	4372	100.00%		TOTAL	517	100.00%

no Brasil[4], a área média por município é superior a 1.500 km^2. Nesse caso, relações espaciais como '(pertinho|próximo|perto) de' são preferíveis, o que explica o maior número de relações *fuzzy*. Além disso, é muito comum que as distâncias entre o centro de uma cidade litorânea e suas praias sejam expressas em quilômetros. Como Belo Horizonte encontra-se distante do mar, essa importante relação espacial métrica é pouco utilizada.

O maior número de relações topológicas em Belo Horizonte pode ser parcialmente explicado pelo fato do termo 'shopping' ter sido, inadvertidamente, desconsiderado na lista de termos genéricos utilizados durante a extração de termos candidatos (passo 3 da Figura 3.1), ao passo que todos os grandes *shopping centers* de Belo Horizonte

[4] `ftp://ftp.ibge.gov.br/Organizacao_do_Territorio/Areas_e_Limites/Areas.zip`

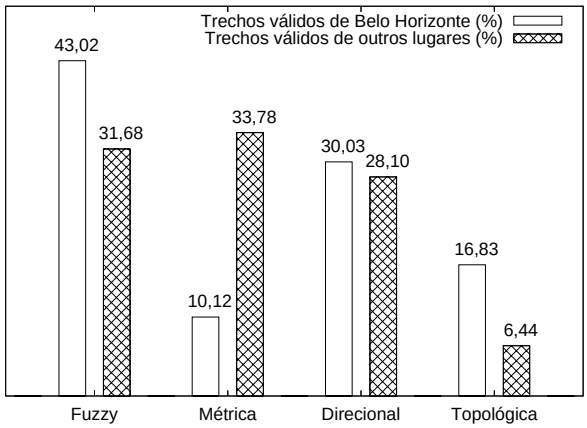

Figura 3.5: Distribuição das relações espaciais por categoria, em Belo Horizonte e nos demais municípios.

Tabela 3.3: Principais tipos de local de interesse em Belo Horizonte.

#	Tipo de Local	Ocorrências	
1	Hospedagem	117	15.73%
2	Placa de publicidade	110	14.78%
3	Outros	38	5.11%
4	Empresas — produtos e serviços diversos	32	4.30%
5	Logradouro	31	4.17%
6	Imóvel	30	4.03%
7	Bar/Café	29	3.90%
8	Restaurante/Lanchonete	27	3.63%
9	Escola	25	3.36%
10	Instalações de empresas, órgãos, repartições e projetos públicos	25	3.36%
	SUB–TOTAL	561	75.40%
	TOTAL	744	100.00%

foram incluídos na lista de pontos de referência utilizados tanto na coleta quanto na extração (passos 1 e 2 da Figura 3.1). Por isso, relações como 'no ? (andar|piso) de' e 'na praça de alimentação de' foram muito mais utilizados em Belo Horizonte. Há ainda algumas diferenças regionais. 'Defronte de', expressão comum no Sul do Brasil, foi pouco utilizada em Belo Horizonte, ao contrário de 'a ? quarteirões de', mais utilizada em Belo Horizonte do que nos demais municípios, possivelmente em função do tamanho regular dos quarteirões na área central da cidade, projetada no final do século XIX.

Em se tratando de referências geográficas, um mesmo local pode aparecer tanto como um local de interesse quanto como um ponto de referência. Entretanto, alguns tipos de local estão claramente mais propensos a figurar como locais de interesse e outros como pontos de referência. Isso é o que nos mostram as Tabelas 3.3 e 3.4 —

Tabela 3.4: Principais tipos de ponto de referência em Belo Horizonte.

#	Tipo de Local	Ocorrências	
1	Centro comercial	129	14.19%
2	Edificação de destaque	106	11.66%
3	Universidade/Faculdade	95	10.45%
4	Divisão territorial	91	10.01%
5	Parque natural/Área de lazer	56	6.16%
6	Praça	56	6.16%
7	Espaço para eventos	37	4.07%
8	Estádio/Ginásio	34	3.74%
9	Hospital	33	3.63%
10	Estação/Terminal rodoviário	32	3.52%
	SUB–TOTAL	788	86.69%
	TOTAL	909	100.00%

nenhum dos principais tipos de local de interesse aparece na tabela dos principais tipos de ponto de referência, e vice–versa. Obviamente existem locais encontrados em ambos os "lados" de uma referência geográfica, mas esses representam uma porcentagem menor do total de locais.

Esse resultado ilustra na prática o que intuitivamente poderíamos esperar devido à nossa familiaridade com o emprego cotidiano de expressões de posicionamento: serviços típicos de páginas amarelas e classificados como hotéis, imóveis, bares, restaurantes e escolas figuram como locais de interesse, tendo suas localizações descritas mediante a utilização de referências a locais bastante conhecidos na cidade, como *shopping centers*, universidades, parques, praças e hospitais. A Tabela 3.5 reforça essa idéia, exibindo os principais pontos de referência de Belo Horizonte: edificações de grande porte ou importantes divisões territoriais cujas localizações são conhecidas pela maioria da população.

Para finalizar esta seção, é importante destacar a surpreendente presença do tipo *Placa de Publicidade* como o segundo tipo de local de interesse mais freqüente, atrás apenas de *Hospedagem*, que engloba, entre outros, hotéis, pousadas e albergues. Documentos relacionados a placas de publicidade contêm uma grande concentração de referências geográficas, o que explicaria esse bom desempenho em nossa coleção, onde existem apenas três documentos com essa característica: para esse tipo de local, as referências por expressões de posicionamento chegaram a uma média de 18,3 por documento, valor alto se comparado ao de outros tipos de local — para hospedagem, por exemplo, esse valor é de cerca de 1,7 referências por documento. Isso indica a importância de se caracterizar bem a localização desse tipo de mídia publicitária como forma de atrair e manter anunciantes.

Tabela 3.5: Principais pontos de referência em Belo Horizonte.

#	Ponto de Referência	Ocorrências	
1	Centro de Belo Horizonte	45	4.95%
2	PUC Minas - Pontifícia Universidade Católica de Minas Gerais - Coração Eucarístico	34	3.74%
3	Região da Savassi	32	3.52%
4	BH Shopping	29	3.19%
5	Minascentro - Centro Mineiro de Promoções Israel Pinheiro	29	3.19%
6	Parque Municipal Américo Renée Giannetti	29	3.19%
7	UFMG - Universidade Federal de Minas Gerais - Campus Pampulha	28	3.08%
8	Aeroporto da Pampulha	27	2.97%
9	Shopping Del Rey	27	2.97%
10	Mineirão - Estádio Governador Magalhães Pinto	24	2.64%
	SUB–TOTAL	304	33.44%
	TOTAL	909	100.00%

3.3 Expandindo os Horizontes

O processo empregado no estudo de caso descrito na Seção 3.2 para identificar expressões de posicionamento em documentos da *Web* baseia–se em dois conjuntos de dados: (1) nomes de pontos de referência; e (2) expressões regulares representando as relações espaciais. Na língua portuguesa, as relações espaciais constituem um conjunto estático e pequeno. Em nossa coleção, a maioria das expressões de posicionamento identificadas contêm o mesmo sub–conjunto de relações espaciais: as dez mais freqüentes ocorrem em cerca de 90% das expressões de posicionamento, enquanto que as vinte seguintes são encontradas nos 10% restantes (Tabela 3.2). Os elementos do conjunto de nomes de pontos de referência foram selecionados com base apenas no conhecimento de Belo Horizonte adquirido pelo autor e seus colegas. Apesar de discutivelmente essa ser uma limitação, é importante observar que a maioria dos envolvidos vive na cidade há décadas; alguns trabalham com o SIG municipal há mais de 15 anos e estão, portanto, familiarizados com os principais pontos de referência da cidade. Entretanto, apesar do esforço para torná-lo o mais completo e representativo possível, os resultados mostraram que estávamos longe de atinger esse objetivo: dos 225 pontos de referência distintos identificados, 91 não pertenciam à lista inicial de nomes e foram encontrados durante o processo de classificação. Além disso, ao contrário do que foi observado com relação à representatividade das principais relações espaciais, os dez pontos de referência mais importantes estão presentes em apenas 33% das expressões de posicionamento. Esse

fato sugere que, à medida que o tamanho do conjunto de pontos de referência cresce, as chances de encontrarmos expressões de posicionamento no texto de um documento aumentam.

A fonte habitual para nomes de lugares é um *gazetteer* (Hill, 2000). Mesmo existindo vários *gazetteers* disponíveis na *Web*, como *ADL Gazetteer*[5], *GNS*[6] e *GNIS*[7], eles geralmente possuem dados completos e atualizados apenas para países desenvolvidos, especialmente os Estados Unidos. No Brasil e em muitos países em desenvolvimento, fontes de dados como essas encontram–se em estágio inicial de criação ou simplesmente não existem. Mesmo os *gazetteers* citados na literatura como fonte de nomes de entidades geográficas do mundo inteiro contêm muito pouca informação. Além disso, a maioria dos lugares refere–se a divisões geo–políticas (áreas administrativas) e a recursos hidrográficos ou hipsográficos, apresentando uma deficiência em termos de dados de locais intra–urbanos, característica típica dos pontos de referência. O *GNS*, por exemplo, contém 87.608 nomes de locais e lugares no Brasil. Esses dados encontram–se desatualizados (98% dos registros foram inseridos ou atualizados entre 1993-1999) e não há distinção entre nomes atuais e antigos (ex.: os nomes 'Território de Guaporé' e 'Território de Rondônia' referem–se ao 'Estado de Rondônia'). O tipo de local intra–urbano mais freqüente, 'escola', possui tão somente 332 registros. Para efeito de comparação, o *GNIS* possui catalogadas, apenas para o estado de Nova York/EUA, 4.961 escolas.

Mesmo se fôssemos capazes de enumerar todos os pontos de referência de Belo Horizonte, ainda assim teríamos problemas, pois o emprego de um ponto de referência em uma expressão de posicionamento está intimamente ligado ao contexto onde ele se insere. Mudanças no ambiente podem interferir na forma como um local é utilizado como um ponto de referência. A paisagem urbana, que serve de contexto para as referências geográficas, muda muito rapidamente, acarretando a necessidade de uma atualização constante do *gazetteer*. Locais antes relevantes e empregados como pontos de referência podem perder a importância ao longo do tempo, passando a ser cada vez menos utilizados como referência. Do mesmo modo, novos locais podem surgir e locais antigos podem receber novas denominações. A incorporação desses nomes de locais na cultura local é o fator que determinará o seu emprego como um ponto de referência.

Para superar essa restrição com relação ao emprego de *gazetteers*, foi implementado um método de identificação de expressões de posicionamento que funciona de forma independente, sem a necessidade de um repositório de nomes de pontos de referência, projetado com base nos dados resumidos na Tabela 3.6. Essa tabela exibe as distâncias

[5] `http://www.alexandria.ucsb.edu/gazetteer`

[6] `http://gnswww.nga.mil/geonames/GNS`

[7] `http://geonames.usgs.gov`

Tabela 3.6: Distância média, em palavras, entre a relação espacial e o ponto de referência, por categoria de relações espaciais.

	Trecho Inválidos	Trechos Válidos
Relações Espaciais	Distância Média	
Fuzzy	3,18	0,28
Métrica	3,00	0,08
Direcionais	2,85	0,07
Topológica	2,61	0,13
Geral	2,93	0,17

médias, em palavras, medidas entre as relações espaciais e os pontos de referência dos trechos válidos e inválidos, assim classificados no estudo de caso de Belo Horizonte da Seção 3.2. Pelos dados da tabela, é fácil perceber que, em uma expressão de posicionamento (trechos válidos), a relação espacial e o ponto de referência encontram–se muito próximos, praticamente não existindo outras palavras entre eles. Já nos trechos inválidos, onde o par ⟨*relação espacial, ponto de referência*⟩ não constitui uma expressão de posicionamento, a distância entre eles é bem maior, quase três palavras na média. Essa característica é válida até mesmo para relações espaciais como 'perto de' e 'dentro de', empregadas em vários outros contextos que não o geográfico — quando aparecem em uma expressão de posicionamento a distância média entre elas e os respectivos pontos de referência é próxima de zero. Dessa forma, podemos assumir que, em uma expressão de posicionamento, o nome de um ponto de referência ocorre logo após a relação espacial.

Com base nas observações acima, o programa extrator de padrões textuais empregado anteriormente no processo de reconhecimento de expressões de posicionamento foi adaptado para trabalhar apenas com um conjunto R de expressões regulares descrevendo as relações espaciais, dispensando o uso de um repositório de nomes de pontos de referência. Abordagem semelhante é empregada no produto comercial *Geographic Text Search (GTS)*, da empresa *Metacarta*[8] para identificar nomes de lugares e em Pasca (2004) para reconhecer categorias e nomes próprios em documentos da *Web*. Na Figura 3.6, um pseudo–algoritmo resume o funcionamento do extrator.

Para cada casamento bem sucedido de um elemento de R em um texto T, uma rotina responsável por identificar um nome de lugar é executada. Essa rotina inicia a pesquisa na posição do texto imediatamente posterior àquela onde a relação espacial foi encontrada. A identificação de nomes baseia–se na ocorrência de letras maiúsculas, empregadas na língua portuguesa para diferenciar nomes próprios, e em outras heurís-

[8] `http://www.metacarta.com`

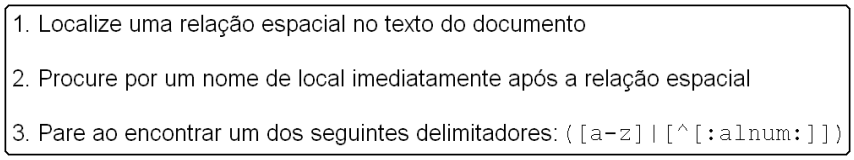

Figura 3.6: Funcionamento resumido do extrator.

ticas de processamento de texto. Sinais de pontuação como '.', ')' e '!', *tags* HTML[9] e palavras em minúsculas delimitam o fim de um nome.

O extrator é capaz de identificar expressões de posicionamento contendo três tipos de nome de local, como ilustrado nos exemplos abaixo, retirados dos documentos da coleção:

1. *nome próprio*: "(...)a duas quadras da **Praça da Liberdade**<>", "(...)ao lado do **Minascentro**.";
2. *nome genérico*: "(...)perto do **centro**.", "(próximo à **prefeitura**)";
3. *nome misto*, uma composição dos tipos acima: "(...)em frente ao **estádio do Mineirão**.", "A 2 km do **aeroporto da Pampulha** existe um(...)";

Além das expressões de posicionamento ilustradas, onde ocorre apenas um ponto de referência, há ainda tratamento para as expressões de posicionamento onde mais de um local é referenciado pela relação espacial, as chamadas *expressões de posicionamento compostas*, como, por exemplo, "(...)perto do **Minascentro** e do **Mercado Central**.".

O extrator foi calibrado utilizando os 909 trechos de Belo Horizonte contendo uma expressão de posicionamento, atingindo uma precisão de quase 99% na identificação correta dos nomes dos pontos de referência, como ilustra a Tabela 3.7. Utilizando o extrator em nossa coleção de documentos da *Web*, foi possível identificar 29.645 expressões de posicionamento, com 13.512 pontos de referência distintos. A qualidade dessa extração foi verificada mediante uma análise por amostragem do conjunto de expressões de posicionamento extraídas. O tamanho da amostra avaliada, cerca de 500 registros, foi determinado por um processo estatístico, de modo que, para um nível de confiança de 95%, é possível afirmar que 89,6 $\pm$ 4,0 % das expressões de posicionamento são válidas. Um valor mais preciso, ou seja, com uma margem de erro menor, pode ser obtido aumentando–se o tamanho da amostra avaliada.

Esse índice atribuído à qualidade da extração, um valor que pode ser considerado satisfatório, pode ser melhorado mediante uma análise mais completa e individualizada do conjunto de expressões de posicionamento extraídas. Identificando–se as principais

[9]Durante o pré–processamento, as *tags* HTML foram contraídas para '<>'.

Tabela 3.7: Desempenho do extrator para as expressões de posicionamento dos trechos de Belo Horizonte. As porcentagens referem–se ao valor do item no nível superior.

Descrição	Ocorrências	
Total	909	100.00%
Ponto de referência *não* encontrado	126	13.86%
Ponto de referência encontrado	783	86.14%
Nome correto	774	98.85%
Nome incorreto	9	1.15%

expressões que aparecem após cada uma das relações espaciais e que são consideradas pelo extrator um local, quando na verdade *não são*, é possível criar uma lista de *stopwords*, expressões a serem desconsideradas pelo extrator, diminuindo a incidência de *falsos positivos*. Para ilustrar, as palavras 'TV', 'Deus', 'Senhor' e 'Estado' aparecem associadas a expressões que denotam relações espaciais com certa freqüência, como em 'perto de Deus', 'em frente à TV', 'no coração do Senhor' e 'dentro do Estado'.

Um dado que chamou atenção foi a quantidade de documentos contendo uma ou mais expressões de posicionamento — 11.485 dos 75.410 documentos, o que corresponde a 15,23% — uma porcentagem expressiva se comparada àquelas encontradas para outros localizadores geográficos, como, por exemplo, códigos postais (4,5%) e números de telefone (8,5%) (McCurley, 2001). Esse resultado pode ser contestado, uma vez que os documentos da coleção foram selecionados justamente por conterem pontos de referência notórios, o que poderia tornar a coleção propensa à existência de expressões de posicionamento. Desse modo, o extrator foi aplicado a uma outra coleção, a `WBR05`, composta por quase 3,6 milhões de documentos coletados de *sites* da *Web* brasileira[10] em março de 2005. Como resultado, 213.093 expressões de posicionamento contendo 51.108 pontos de referência distintos foram extraídos. No caso da `WBR05`, 3,6% dos documentos possuem uma ou mais expressões de posicionamento, com uma média de 1,6 expressões por documento. Pode–se perceber, portanto, que a quantidade de expressões de posicionamento nos documentos de Belo Horizonte, bem acima do valor encontrado para a `WBR05`, deve–se à forma com que a coleção foi obtida, tornando–a propensa a apresentar esse tipo de construção. Não obstante, o valor encontrado é compatível com o de outras fontes de contexto geográfico, o que confirma a importância das expressões de posicionamento.

[10] *Sites* com o código de país 'br' (Brasil) no nome do domínio, como definido pela norma *ISO 3166-1 alpha-2*

3.4 Interpretação das Expressões de Posicionamento

Após a discussão, na seção anterior, sobre como identificar expressões de posicionamento em documentos textuais e tendo implementado uma ferramenta para fazê-lo, é necessário desenvolver um método para determinar de forma quantitativa a posição no espaço que elas descrevem qualitativamente. Com isso, em aplicações de busca local, poderemos, por exemplo, verificar se o contexto geográfico atribuído por elas a um documento é compatível com o escopo geográfico de uma consulta.

A interpretação de uma expressão de posicionamento em linguagem natural pode ser um assunto complexo, pois a percepção da dimensão por ela representada está intimamente ligada a fatores subjetivos e dependentes do contexto. A escala, por exemplo, é um fator importante. Compare as seguintes referências geográficas: $r1$ = ('Ouro Preto', 'perto de Belo Horizonte') e $r2$ = ('Detran', 'perto da Praça da Liberdade'). Em $r1$, 'perto de' representa uma distância de aproximadamente 100 quilômetros, enquanto que, em $r2$, representa cerca de 500 metros. Assim, dada uma expressão de posicionamento com a relação espacial 'perto de', qual o valor, a medida que se pode atribuir a ela? Neste trabalho, nos restringiremos a locais intra–urbanos, o que faz diminuir o impacto da escala. Entretanto, outros fatores devem ser observados.

Worboys (2001) descreve um experimento para tentar capturar a noção que as pessoas possuem da proximidade entre locais. Através de um questionário, um grupo de pessoas indicou, para cada local em uma lista, se ele está próximo a um outro local, designado como ponto de referência. Outro grupo indicou, para os mesmos locais e pontos de referência, aqueles que *não* estão próximos. Os locais encontram–se em um *campus* universitário. Pelos resultados, percebe–se que alguns locais foram indicados como estando ao mesmo tempo próximos e não–próximos de um mesmo ponto de referência. Posteriormente, essa experimentação foi ampliada para avaliar a noção de direção (Worboys et al., 2004). Outro trabalho relevante, na mesma linha do anterior, é o de Montello et al. (2003). Nele, um estudo empírico é conduzido para determinar os limites de uma região, no caso o centro da cidade de Santa Barbara/CA, com base na crença do que as pessoas pensam ser tal região. Por esses experimentos, pode–se perceber que mesmo sem entrar no mérito quantitativo, as pessoas podem interpretar um conceito vago, como uma região ou relação espacial, de formas diferentes e até mesmo contraditórias.

Egenhofer e Shariff (1998) derivam métricas do modelo topológico *9–intersection* (Egenhofer et al., 1994) capazes de determinar, entre 64 relações espaciais em inglês, a mais apropriada para descrever uma configuração entre dois objetos geométricos. Parece, entretanto, não haver qualquer tipo de regra que nos leve a empregar, em uma expressão de posicionamento, certas relações espaciais em detrimento de outras. Não

Tabela 3.8: Distâncias *em metros* correspondentes às relações espaciais.

Relação Espacial	MÉD	MÍN	MÁX
dentro de	0,00	0,00	0,00
ao lado de	142,00	9,30	419,30
em frente a	183,40	27,00	480,90
atrás de	346,60	105,90	794,90
perto de	527,00	99,20	2.159,90
a 1 minuto de	530,90	109,70	2.507,70
próximo a	634,50	26,10	3.433,70
nas proximidades de	782,00	277,60	1.882,10
a 1 km de	1.000,00	1.000,00	1.000,00
a 1 quilômetro de	1.000,00	1.000,00	1.000,00

há uma noção formal ou restrição de linguagem (pelo menos em Português), que faça com que uma relação espacial seja mais adequada para determinadas situações. Pelo contrário, sugere que possa haver uma *equivalência* entre expressões de posicionamento gerada pela sobreposição de significados de relações espaciais. Por exemplo, dois locais podem ser definidos como estando 'próximos' ou 'perto' um do outro, o que não invalida o fato deles também poderem estar a 'X metros', 'Y quilômetros' ou 'Z minutos' de distância.

Para tentar identificar essas equivalências entre expressões de posicionamento, foi realizado um experimento para determinar a distância média entre os locais que as referências espaciais encontradas em documentos da *Web* tentam representar. Dessa forma, espera–se obter um valor médio para cada uma das relações espaciais. O experimento está descrito a seguir:

1. Os endereços dos pontos de referência das expressões de posicionamento dos 555 trechos classificados no estudo de caso de Belo Horizonte, segundo a Tabela 3.1, como sendo 'Trecho válido, em Belo Horizonte, cujo objeto de interesse é um local com endereço', juntamente com os endereços dos locais de interesse aos quais eles se referem, passaram por um processo de geocodificação. Esse processo foi realizado com auxílio dos dados e ferramentas do SIG da cidade de Belo Horizonte, sendo que, ao final, foi possível obter 418 referências geográficas geocodificadas, compostas por 257 locais de interesse, 105 pontos de referência e 23 relações espaciais distintas. Locais cuja área é significativa, como a Lagoa da Pampulha ou o Estádio do Mineirão, tiveram sua localização aproximada pelas coordenadas do centróide de seu *Retângulo Mínimo Envolvente (RME)*.

2. As referências geocodificadas foram agrupadas em 23 conjuntos, cada um representando uma relação espacial. Dentro de cada conjunto, as distâncias Eu-

clidianas entre as coordenadas geográficas do par ⟨*local de interesse, ponto de referência*⟩ foram calculadas, assim como as distâncias mínimas, médias e máximas dentro de cada conjunto. A Tabela 3.8 exibe os valores encontrados para as principais relações espaciais (as mais freqüentes). Nesse experimento, não calculamos as distâncias para as relações topológicas e nem para as seguintes relações métricas: 'a ? (km|quilômetros|metros|m) de'. No primeiro caso, atribuímos um valor de distância igual a zero, pois consideramos que o local de interesse possui as mesmas coordenadas geográficas do ponto de referência. No segundo, tomamos como distância o valor nominal informado pela relação espacial, ou seja, 'a 10 km de' corresponde a uma distância de 10.000 metros entre o local de interesse e o ponto de referência.

Pelos dados da Tabela 3.8, pode–se perceber que muitas relações espaciais com interpretações semânticas diferentes, mesmo opostas, como é o caso das relações 'em frente de' e 'atrás de' são equivalentes no que se refere à distância entre o local de interesse e o ponto de referência. De fato, com exceção das relações métricas, em especial aquelas expressas em minutos ou quilômetros, todas as demais relações podem ser consideradas equivalentes em termos de distância. Para ilustrar esse novo conceito, a Figura 3.7 exibe a localização de alguns locais de interesse em relação a dois pontos de referência em Belo Horizonte, o Mercado Central e o Diamond Mall. Apesar de diferentes relações espaciais terem sido usadas na expressões de posicionamento para descrever a posição relativa dos locais de interesse, quase todos encontram–se dentro do limite determinado pela relação espacial 'próximo a'.

Cabe ressaltar que os valores encontrados para as relações espaciais obviamente valem para a cidade de Belo Horizonte, e apenas se considerarmos as expressões de posicionamento avaliadas como sendo um conjunto representativo das expressões de posicionamento utilizadas para descrever a posição de locais de interesse nessa cidade. Não obstante, eles fornecem uma boa idéia a respeito do que o *senso comum* entende por expressões como 'perto de' e 'próximo de'. Em outros lugares, dependendo de fatores como o tamanho do município, fluidez do trânsito, tamanho médio dos quarteirões, etc., esses valores podem apresentar variações. Dessa forma, quanto mais representativo o conjunto de expressões de posicionamento utilizadas para estimar as distâncias expressas pelas relações espaciais, mais precisos serão esses valores. Para incluir mais expressões de posicionamento nesse cálculo, é necessário, porém, como fizemos no experimento descrito nesta seção, realizar as seguintes operações:

1. Localizar expressões de posicionamento no texto dos documentos;

2. Localizar no texto dos documentos os locais de interesse aos quais as expressões de posicionamento se referem;

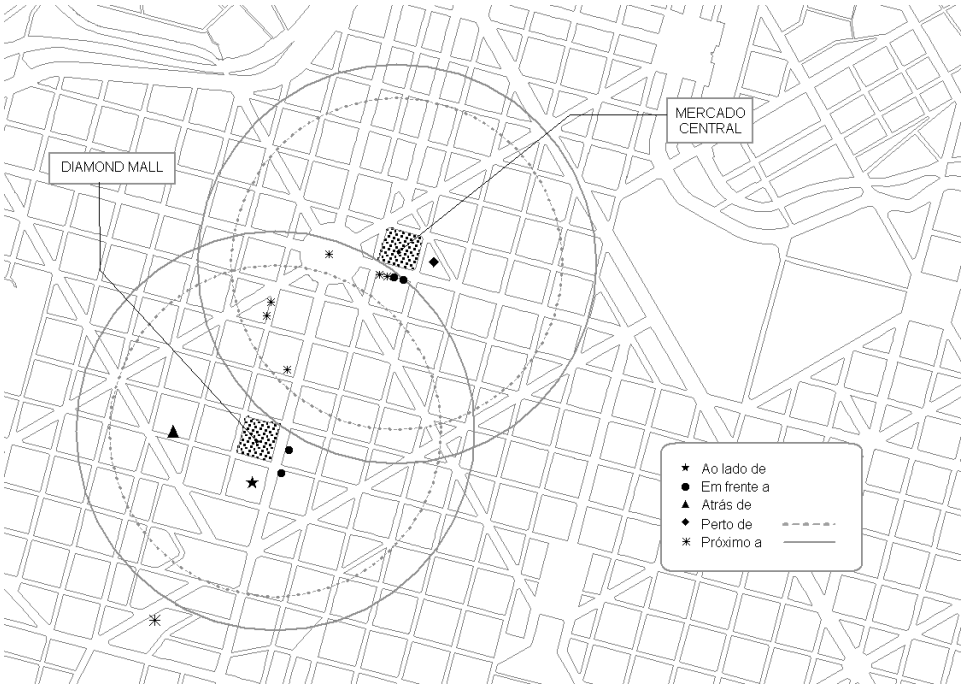

Figura 3.7: Posição dos locais de interesse em relação à área determinada pelas relações espaciais 'perto de' e 'próximo a'.

3. Retirar possíveis ambigüidades presentes nos nomes dos locais de interesse e pontos de referência;

4. Geocodificar os locais de interesse e os pontos de referência encontrados.

O item (1) pode ser feito com ajuda do extrator descrito na Seção 3.3. Já os itens de (2) a (4) fogem ao escopo deste trabalho, pertencendo ao conjunto de desenvolvimentos futuros. Mesmo o processo podendo ser considerado incompleto, é possível projetar várias aplicações de RIG que utilizam as expressões de posicionamento e os valores de distância encontrados para as relações espaciais, como mostra o capítulo seguinte.

Capítulo 4

Aplicações

Neste capítulo, descrevemos na Seção 4.1 uma aplicação de busca local que utiliza as expressões de posicionamento presentes em documentos da *Web* para recuperar documentos contendo possíveis informações de interesse relacionadas a um local próximo a um ponto de referência (Delboni et al., 2005). Na seção seguinte, introduzimos uma estrutura denominada *Grafo de Inferência Geográfica* que permite que as informações de cunho geográfico transmitidas pelas expressões de posicionamento possam ser armazenadas e recuperadas. Além disso, é possível definir operações sobre essa estrutura de modo a realizar alguns tipos de inferência, como a determinação de locais próximos, sem que para isso seja necessário geocodificar a posição no espaço de locais de interesse e pontos de referência.

4.1 Uso de Expressões de Posicionamento em Busca Local

Dada uma consulta $Q = (A, D, P)$ formada por um conjunto de palavras–chave A representando um objeto de interesse (por exemplo, $A = \{$'hotel', 'de', 'luxo'$\}$), uma distância D (por exemplo, $D =$ '2 km') e uma seqüencia de caracteres P contendo o nome de um ponto de referência (por exemplo, $P =$ 'praia de Copacabana'), o *objetivo* da aplicação aqui descrita é recuperar documentos onde ocorra A e exista pelo menos uma expressão de posicionamento onde o ponto de referência é P e cuja posição no espaço, determinada pela relação espacial que o acompanha, esteja dentro do escopo definido por D em relação a P, isto é, a uma distância máxima D de P. Para a consulta Q exemplificada acima, isso equivaleria a recuperar documentos sobre 'hotéis de luxo localizados a no máximo 2 km da praia de Copacabana'. Se essa consulta fosse utilizada em uma máquina de busca convencional, poderíamos obter resultados incompletos (documentos com expressões de posicionamento contendo outras relações

espaciais que não 'a 2 km de' ficam de fora) e imprecisos (documentos onde as palavras 'praia' e 'copacabana' não são usadas no contexto geográfico).

Em nossa aplicação, o ponto de referência é utilizado como centro de busca, de forma análoga aos demais localizadores geográficos nas atuais aplicações de busca local, como o *Google Local*. O centro de busca, juntamente com o raio de busca, determina uma área na superfície terrestre com a qual objeto de interesse procurado pelo usuário deve estar relacionado. Essa área é denominada de *escopo geográfico da consulta*. Nas aplicações de busca local, o processamento de uma consulta com escopo geográfico, em linhas gerais, pode ser descrito em três passos: (1) o centro de busca é geocodificado, atribuindo–se a ele uma coordenada geográfica; (2) um conjunto de documentos contendo indicadores geográficos, previamente localizados, geocodificados e indexados, compatíveis com o escopo geográfico da consulta é selecionado; e (3) um algoritmo tradicional de recuperação de informação é executado para ordenar os documentos com base nas palavras–chave utilizadas para especificar o objeto de interesse e em outras informações importantes, como os *links* que apontam para os documentos.

Existem várias diferenças entre a abordagem ilustrada nesta seção e as outras aplicações de busca local: (1) durante a fase de indexação, não é necessário determinar as coordenadas dos localizadores geográficos encontrados no texto dos documentos. Isso elimina os procedimentos utilizados para resolver a ambigüidade de nomes, porém confia ao algoritmo de *ranking* essa tarefa; (2) pontos de referência, um conceito familiar para as pessoas, são utilizados como centro de busca; e (3) a compatibilidade geográfica entre o documento e o escopo geográfico da consulta é baseada apenas no fato dele conter uma expressão de posicionamento em acordo com o escopo determinado, e não em cálculos envolvendo coordenadas geográficas. O algoritmo de *ranking* é comum às duas abordagens.

Como pode–se perceber, o tipo de processamento envolvido é muito simples, podendo ser implementado de forma eficiente. O ponto–chave é determinar expressões de posicionamento que se enquadrem no escopo geográfico da consulta, ou seja, que sejam *compatíveis*. Por exemplo, dado o escopo 'a no máximo 3 quilômetros da Praça da Liberdade', quais são as expressões de posicionamento compatíveis? Como veremos a seguir, isso é feito com ajuda dos valores médios de distância calculados para as relações espaciais no Capítulo 3 e que constam na Tabela 3.8.

A Seção 4.1.1 descreve uma maneira de implementar essa estratégia em uma máquina de busca convencional, enquanto que a Seção 4.1.2 mostra uma aplicação prática que implementa uma meta–busca local envolvendo expressões de posicionamento utilizando o *Google*.

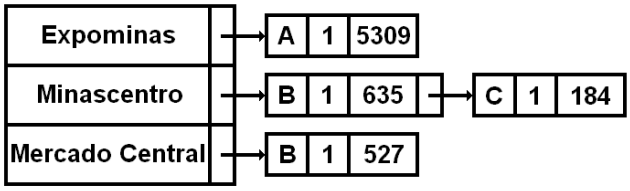

Figura 4.1: Exemplo de uma estrutura parcial representando uma lista invertida.

4.1.1 Visão Geral da Estratégia de Busca

Nossa estratégia de busca baseia–se na expansão de consultas e pode ser explorada por virtualmente qualquer máquina de busca convencional, que funcione por palavras–chave. Ela se apóia em pequenas modificações nas fases de indexação e pesquisa de documentos e requer intervenções pouco significativas nas estruturas de dados internas, porém deixa o algoritmo de *ranking* intocado.

Durante a fase de indexação, o *parser* (analisador sintático) deve ser modificado para incluir uma rotina de identificação de expressões de posicionamento. Tal rotina, como observado na Seção 3.3, pode ser implementada por um programa de extração de informações, como o GATE (Cunningham et al., 2002), que utiliza uma série de regras para casamentos de cadeias de caracteres para identificar *tokens* de interesse e extrair a informação desejada, no caso, as expressões de posicionamento.

A lista de expressões de posicionamento obtida para um dado documento precisa então ser processada. Em cada expressão de posicionamento as palavras compondo o nome do ponto de referência devem ser tratadas como um único termo e inseridas na lista invertida da máquina de busca, da mesma forma que qualquer outro termo encontrado no corpo de um documento seria. Opcionalmente, poderia ser inserido em uma lista invertida separada, dedicada apenas a pontos de referência. A relação espacial deve ser convertida para um valor quantitativo, como a distância média em metros que ela representa, e armazenado como uma informação adicional no registro correspondente à ocorrência do ponto de referência ao qual ela se refere. A esse valor, daremos o nome de *distância*.

A Figura 4.1 exibe o estado de uma lista invertida imaginária logo após o processamento das expressões de posicionamento 'a 10 minutos do Expominas' no documento A, 'próximo ao Minascentro' e 'perto do Mercado Central' no documento B, e 'em frente ao Minascentro' no documento C. O registro para cada ocorrência de um ponto de referência é representado por uma tripla contendo o identificador do documento, a freqüência do termo (nome do ponto de referência) no documento, e a distância, nesse caso em metros, expressa pela relação espacial. O último valor foi obtido da

Tabela 3.8. No caso da relação espacial métrica 'a 10 minutos de', o valor em metros para 'a 1 minuto de' foi multiplicado por 10.

Como visto na Seção 3.3, as expressões de posicionamento ocorrem em cerca de 3,6% dos documentos da `WBR05`, com uma média de 1,6 expressões por documento. Com base nesse padrão de ocorrência é possível dizer, grosso modo, que os pontos de referência serão encontrados em uma quantidade duas ordens de grandeza menor do que as demais palavras–chave. Como consequência, podemos esperar um impacto muito pequeno no tamanho e no processamento de uma lista invertida que armazene pontos de referência.

As adaptações necessárias na fase de processamento de consultas podem ser divididas em dois passos. Primeiramente, a consulta do usuário deve ser avaliada, para tentar encontrar um dos nomes de pontos de referência existentes na lista invertida. A avaliação pode ser feita, no caso de uma máquina de busca convencional, pelo mesmo módulo responsável por produzir recomendações, como correções ortográficas e exibição de *links* patrocinados, a partir de uma análise sintática da consulta. Nesse caso, para a consulta 'hotel aeroporto da pampulha', o módulo poderia retornar o seguinte: '*Procurando por* ***hotel*** *próximo ao* ***Aeroporto da Pampulha****?*'. O ponto de referência poderia ainda ser informado em separado, o que restringiria a análise apenas às palavras digitadas na caixa de texto reservada a esse fim. Opcionalmente, o significado semântico em termos de distância determinado pela relação espacial 'próximo ao' na consulta sugerida, poderia ser especificado pelo usuário, que escolheria uma distância adequada à sua necessidade de informação (ex: 800 m, 5 km, etc.) ou um valor padrão pode ser atribuído pelo sistema, como, por exemplo, '3 km'. Esse valor, denominado *raio de busca*, juntamente com o ponto de referência definem o escopo geográfico da consulta.

Em seguida, os documentos contendo as palavras–chave da consulta devem ser selecionados. Portanto, para a consulta do exemplo acima, deve–se selecionar todos os documentos contendo as palavras–chave 'hotel' e 'aeroporto da pampulha', essa última identificada como um ponto de referência e portanto tratada como um único termo. Além disso, existe uma restrição adicional que deve ser respeitada: o escopo geográfico da consulta, definido pela expressão de posicionamento 'próximo ao aeroporto da pampulha', deve ser compatível com as expressões de posicionamento encontradas no texto do documento cujo ponto de referência é o 'aeroporto da pampulha'. Para isso, comparamos o valor do raio de busca (ex.: 3.000 metros) com a informação de *distância* armazenada na lista invertida de termos nos registros de cada ocorrência do ponto de referência 'aeroporto da pampulha' no texto dos documentos. Dessa forma, é possível recuperar documentos contendo expressões de posicionamento sintaticamente diferentes, porém compatíveis em termos do significado semântico da distância fornecido pela

Figura 4.2: Aplicação de busca na *Web*

interpretação das relações espaciais — como se estivéssemos selecionando sinônimos para uma expressão de posicionamento.

Para concluir a etapa de processamento, os documentos selecionados são ordenados e exibidos ao usuário, ou seja, seguem o fluxo normal de uma máquina de busca. Opcionalmente, o algoritmo de *ranking* pode ser modificado para incluir o valor da distância nos cálculos de pesos dos documentos. Dessa forma, se os documentos d_1 e d_2 possuírem pesos semelhantes, mas a distância do ponto de referência em questão corresponde a 3.342 metros no documento d_1 e 589 metros no documento d_2, d_2 poderia ser melhor classificado do que d_1.

4.1.2 Resultados Experimentais

Para avaliar a plausibilidade da estratégia de busca local proposta, foi realizado um pequeno experimento para simulá-la utilizando uma máquina de busca comercial como processador de consultas. Para realizar esse experimento, a interface de busca exibida na Figura 4.2 foi implementada. Ela recebe como entrada o conjunto A de palavras–chave que especifica um objeto de interesse e o nome P de um ponto de referência. Para tentar recuperar documentos relevantes para essa consulta, P é concatenado junto a várias relações espaciais em linguagem natural, formando diferentes expressões de posicionamento. As expressões de posicionamento são então relacionadas por um operador

Tabela 4.1: Consultas originais utilizadas no experimento

#	Consulta	Cidade
1	hotel aeroporto da pampulha	BHZ
2	hotel aeroporto de congonhas	SAO
3	hotel aeroporto santos dumont	RIO
4	hotel–fazenda próximo a belo horizonte	BHZ
5	hotel congresso nacional	BSB
6	hotel metrô consolação	SAO
7	hotel minascentro	BHZ
8	hotel morumbi shopping	SAO
9	hotel praça tiradentes	OUR
10	hotel riocentro	RIO

booleano OR, formando uma expressão lógica, que é então relacionada ao conjunto A por meio de um operador booleano AND. A expressão resultante é formatada de modo adequado e submetida como uma consulta para a máquina de busca, que irá processar a consulta e retornar uma lista de documentos ordenada por relevância.

Um exemplo ajudará a compreender esse processo. Suponha que uma pessoa, em viagem a Nova York, esteja interessada em encontrar teatros próximos ao Central Park. Usando nossa interface, ele poderá representar essa necessidade de informação especificando A={'teatro'} e P='central park'. Ao submeter a consulta, ela é expandida para a seguinte expressão: "teatro AND ('RE_1 central park' OR 'RE_2 central park' OR ... OR 'RE_n central park')". Palavras entre aspas simples formam frases, devendo aparecer no documento exatamente da mesma forma, e $RE_1...RE_n$ são relações espaciais em linguagem natural como 'minutos de' e 'nas proximidades de', utilizadas no intuito de recuperar documentos onde diferentes relações espaciais acompanham o mesmo ponto de referência. Essa consulta é enviada para um máquina de busca, no caso desse experimento, o *Google*, eleita por prover uma sintaxe de consultas avançadas adequada às nossas necessidades, apesar de restringir a 32 o número total de termos na consulta.

Para determinar se nossa abordagem pode trazer ganhos para uma máquina de busca, aumentando a qualidade dos documentos retornados, comparamos a performance de 10 consultas expressas de modo convencional, como um conjunto de palavras–chave, em relação às suas versões expandidas. A Tabela 4.1 mostra as consultas originais avaliadas.

A escolha de 'hotel' ('hotel–fazenda' para a consulta 4) como objeto de interesse foi baseada em uma análise de um *log* de consultas, compreendendo um período de seis meses[1], do TodoBR[2], uma máquina de busca voltada para a *Web* brasileira, re-

[1] De outubro de 2004 a março de 2005

[2] http://www.todobr.com.br

centemente adquirida pela *Google Inc.*: a análise revelou que as consultas relacionadas a 'hotel' eram pelo menos quatro vezes mais freqüentes que as relacionadas a outros lugares procurados, como restaurantes, teatros, bares, museus e cinemas.

Nas consultas originais, com exceção da consulta 4, palavras–chave denotando relações espaciais foram suprimidas. Pela análise do *log* de consultas, percebemos que a presença das relação espacial denotando proximidade é insignificante; apenas a relação espacial topológica 'em', e suas variações 'no' e 'na', aparecem com certa freqüência, e quase sempre junto a lugares, como 'em Belo Horizonte'. Os trabalhos de Kohler (2003) e Sanderson e Kohler (2004) corroboram essa constatação. Entretanto, a avaliação de consultas presentes no *log* como 'hotel minascentro' ou 'hotel savassi' exige uma análise mais profunda da necessidade de informação expressa por elas. O usuário poderia estar procurando por um hotel específico ('Hotel Savassi'), por hotéis próximos ao lugar/local mencionado ('hotel perto do Minascentro') ou na região delimitada pelo lugar ('hotel na savassi'). Dessa forma, não há como determinar se consultas por locais próximos a um ponto de referência não são relevantes, ou se são e os usuários apenas não empregam as relações espaciais como parte da consulta, ou ainda se as consultas desse tipo não são utilizadas devido à forma como a interface de busca é apresentada ao usuário. Na consulta 4, porém, a expressão 'próximo a' faz parte da consulta. Resolvemos incluí–la pois o conjunto de documentos relevantes deveria conter informações a respeito de hotéis–fazenda nas proximidades de Belo Horizonte, e não 'em Belo Horizonte'. Utilizamos a relação espacial 'próxima a' por ser a mais freqüente segundo o estudo de caso apresentado na Seção 3.2. Ainda com relação a esta consulta, ela é a única a apresentar o nome de um lugar ao invés do nome de um local intra–urbano. Ela foi incluída no intuito de avaliarmos o comportamento da estratégia de expansão de consultas para pontos de referência de grandes dimensões. Com isso, foi possível perceber que as relações espaciais em linguagem natural se adaptam automaticamente ao contexto do ponto de referência que as acompanha. Porém, a conversão das mesmas para valores quantitativos, pra fins de utilização do índice espacial, fica prejudicada. As expressões de posicionamento 'perto de BH', 'próximo à BH' e 'ao lado de BH', por exemplo, são semanticamente equivalentes, mas a conversão para valores quantitativos não pode ser feita segundo a Tabela 3.8 — como o ponto de referência é uma cidade, valores de distância de algumas dezenas de quilômetros passam a ser aceitáveis. Essa é uma das razões pela qual procuramos, neste trabalho, enfatizar apenas locais intra-urbanos como pontos de referência.

Os pontos de referência são importantes locais nos três maiores centros metropolitanos do Brasil — São Paulo (SAO), Rio de Janeiro (RIO) e Belo Horizonte (BHZ) — além da capital Brasília (BSB) e da histórica cidade de Ouro Preto (OUR). Eles foram escolhidos de uma lista de pontos de referência compatíveis com as seguintes

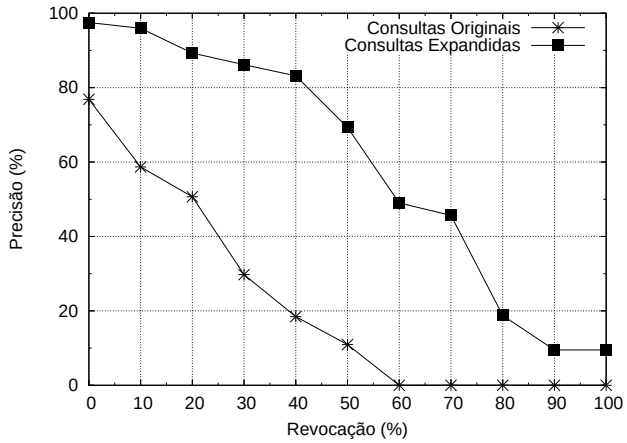

Figura 4.3: Revocação × precisão média para as consultas originais e expandidas

restrições: (1) o nome deve ser formado por, no máximo, duas palavras (os nomes dos aeroportos são uma exceção, como explicado em seguida); (2) o conjunto de resposta retornado para a versão expandida da consulta deve conter, no mínimo, 30 documentos, ou seja, deve ser um ponto de referência representativo; e (3) não pode ser um ponto de referência genérico, como 'prefeitura' ou 'mercado'.

Para avaliar os documentos retornados como resposta às consultas, os 20 primeiros documentos retornados para a consulta original e os 20 primeiros documentos da resposta para a consulta expandida foram combinados em um *pool*, em uma ordem aleatória. Os documentos foram então classificados como sendo relevantes ou não–relevantes por duas pessoas recrutadas voluntariamente em nosso grupo de pesquisa. As diretrizes para a classificação foram dadas da seguinte forma: "Classifique como *relevante* os documentos contendo informação de contato (ex.: número de telefone, endereço, *e–mail* ou *link* para o *site* oficial) de qualquer hotel dito estar *próximo* ao ponto de referência especificado — para as relações espaciais métricas expressas em minutos ou quilômetros, estar *próximo* é estar a no máximo 20 minutos ou 11 km. Todas as demais relações espaciais, direcionais ou *fuzzy*, podem ser consideradas equivalentes a *próximo*. Os demais documentos não são relevantes. OBS: para a consulta 4, estar perto em termos métricos é estar no máximo a 100 km ou 90 minutos".

A Figura 4.3 exibe as curvas de *revocação* × *precisão* médias para os 20 primeiros documentos retornados pelas consultas originais e pelas consultas expandidas. Está claro que os documentos recuperados pelas consultas expandidas são muito mais relevantes, superando a performance das consultas originais em todos os níveis de revocação, com uma excelente precisão média.

Devido ao limite de 32 palavras por consulta imposto pelo *Google*, foi possível utilizar apenas sete relações espaciais diferentes na composição da consulta expandida;

caso contrário, seria necessário dividir a consulta em sub–consultas. Foi justamente isso o que fizemos para as consultas onde os pontos de referência eram aeroportos, pois possuem três palavras compondo seus nomes e são pontos de referência muito importantes (segundo a quantidade de documentos recuperados). Para eles, usamos onze relações espaciais divididas em duas consultas. O defeito dessa divisão da consulta é que, como não temos controle sobre o algoritmo de *ranking*, não é possível saber qual resposta seria gerada pela máquina de busca caso ela aceitasse uma consulta mais longa.

As relações espaciais empregadas foram as que constatamos ser mais freqüentes na língua portuguesa, segundo o estudo de caso da Seção 3.2 e retirando–se as relações espaciais topológicas, já que estamos lidando com locais, e não com lugares. São elas: 'próximo a', 'em frente a', 'a ? km de', 'ao lado de', 'perto de', 'a ? minutos de', 'a ? quilômetros de', 'atrás de', 'nas proximidades de', 'a ? metros de ' e 'a ? quadras de'.

Mesmo que avaliada para um único objeto de interesse e para um número reduzido de pontos de referência, o resultado obtido neste experimento reforça a idéia de que as expressões de posicionamento são importantes fontes de contexto geográfico, utilizadas em conjunto com outros indicadores geográficos para descrever a localização de um determinado objeto de interesse. De fato, quando um nome de local aparece precedido por uma relação espacial no texto de um documento, isso é uma evidência explícita de que algo está tendo a sua localização descrita, ao contrário de quando um nome de local aparece "sozinho" em meio ao texto.

Outro ponto importante ilustrado por essa aplicação é a utilização de um ponto de referência como centro de busca. Considerando que o conceito de ponto de referência é bastante familiar para as pessoas e que a localização por pontos de referência é um método de navegação bastante empregado pelos seres vivos (Moratz e Wallgrün, 2003), é estranho verificar que o uso de pontos de referência como parte de uma consulta geográfica ocorre em poucas aplicações *Web*. Um exemplo é o *site* da *Accor Hotels*[3] que permite localizar hotéis situados próximos a importantes pontos de referência de uma cidade, agrupados em diversas categorias como empresas, compras, aeroportos, pontos turísticos, parques e etc. Outro exemplo é dado pela *Citysearch.com*, que oferece um serviço de busca em páginas amarelas. Na busca por hotéis, pode–se selecionar como parâmetro uma atração turística da cidade (Figura 4.4).

[3] `http://www.accorhotels.com.br`

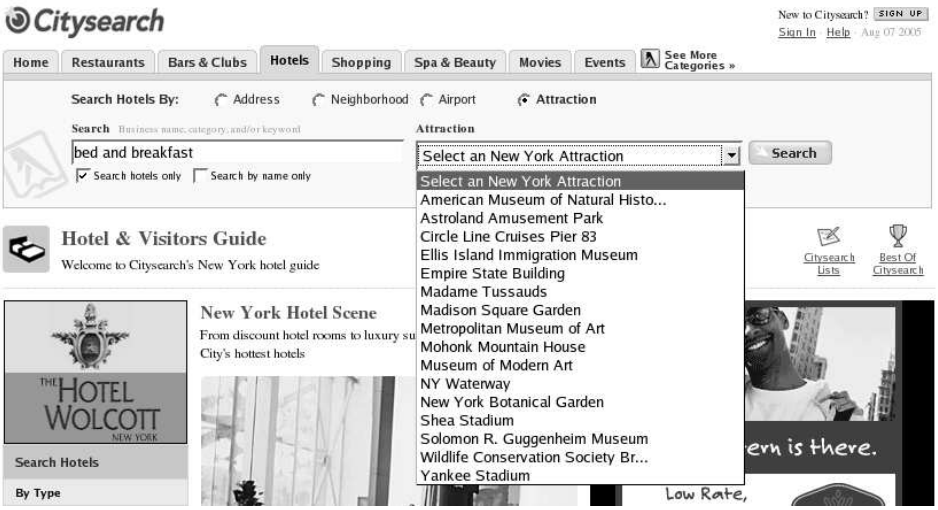

Figura 4.4: Busca por hotéis no *Citysearch.com* utilizando–se ponto de referência

4.2 Grafo de Inferência Geográfica

Em uma parcela de documentos da *Web* ocorrem expressões de posicionamento *compostas*, isto é, relações espaciais acompanhadas por dois ou mais pontos de referência, como 'perto de pr_1, pr_2 e pr_3'. Em outros, existe um conjunto de expressões de posicionamento em seqüência, para tentar descrever com maior precisão a posição de um local em relação a diversos pontos de referência. O trecho a seguir, retirado de um dos documentos da coleção de teste, descreve a posição de um hotel em São Paulo usando várias expressões de posicionamento: *"Próximo à Marg. Tietê, ao lado do Shopping Center Norte. Próximo ao metrô e Terminal Rodoviário Tietê, e ao Campo de Marte (local destinado a taxis aéreos, aviões particulares e helicópteros) 400m - Expo Center Norte 1km - Anhembi 3km - Mart Center Na marginal Tietê, próximo à saída para as rodovias Fernão Dias, Ayrton Senna, Dutra, Anhanguera, Bandeirantes, Castelo Branco, bem como ao Centro Industrial de Guarulhos".*

Aproveitando esse fato, vislumbramos uma forma de determinar a distância aproximada entre dois locais sem a necessidade de identificar locais de interesse, resolver ambigüidade de nomes e geocodificar a posição dos locais. A idéia é armazenar as informações provenientes das referências geográficas em uma estrutura de dados que permita, a partir de construções como 'A está perto de B e C' ou 'D está a 2 km de E e a 3 Km de F', chegar a conclusões do tipo 'B pode estar perto de C' e 'E está a no máximo 5 km de F'. Para isso, vamos introduzir a definição do que denominamos *grafo de inferência geográfica.*

Definição 1 *Um grafo de inferência geográfica G é um grafo, possivelmente desconectado, onde os vértices são pontos de referência ou locais de interesse, e as arestas*

correspondem às relações espaciais entre eles. As arestas são ponderadas e o peso é dado pela distância média correspondente à relação espacial. Em G, todos os vértices possuem grau igual ou maior a um, isto é, não existem vértices isolados. Além disso, cada local distinto é representado por um único vértice.

Pela definição, temos que cada local distinto é representado por um único vértice. Porém, como a princípio não realizamos qualquer tipo de resolução de ambigüidade, pontos de referência com o mesmo nome são representados pelo mesmo vértice, mesmo que se trate de locais distintos. De maneira semelhante, como os locais de interesse associados às expressões de posicionamento extraídas dos documentos não estão sendo identificados, eles são tratados como locais distintos, ou seja, em G, cada local de interesse corresponde a um único vértice, mesmo que eventualmente dois ou mais desses vértices possam se referir a um mesmo local.

Após essas observações, podemos descrever como um grafo de inferência pode ser construído a partir de um conjunto de referências geográficas. Cada referência geográfica é formada por um local de interesse e uma expressão de posicionamento, simples ou composta, ou um conjunto de expressões de posicionamento em seqüência, como exemplificado no início da seção. Assim, dado um grafo G e uma referência geográfica r, a inserção de r em G é feita da seguinte forma:

1. Insira em G um novo vértice u representando o local de interesse.
2. Para cada expressão de posicionamento da referência geográfica, verifique se já existe um vértice v com o nome do ponto de referência.
   a) Se não existir, crie o vértice e adicione uma aresta uv, com o peso w fornecido pela distância média correspondente à relação espacial.
   b) Se v já existir, apenas adicione a aresta uv com peso w.

Para ilustrar esse processo, vamos criar um grafo de inferência para as seguintes referências geográficas:

- $r_1 = \langle li_1, \{(\alpha, pr_1)\}\rangle$
- $r_2 = \langle li_2, \{(\beta, pr_2), (\beta, pr_3)\}\rangle$
- $r_3 = \langle li_3, \{(\beta, pr_1), (\gamma, pr_3)\}\rangle$

Nas referências geográficas acima, li_j é um local de interesse, pr_j é um ponto de referência e as letras gregas são relações espaciais. Note, portanto, que as expressões de posicionamento correspondem às expressões entre parênteses. A Figura 4.5 exibe a

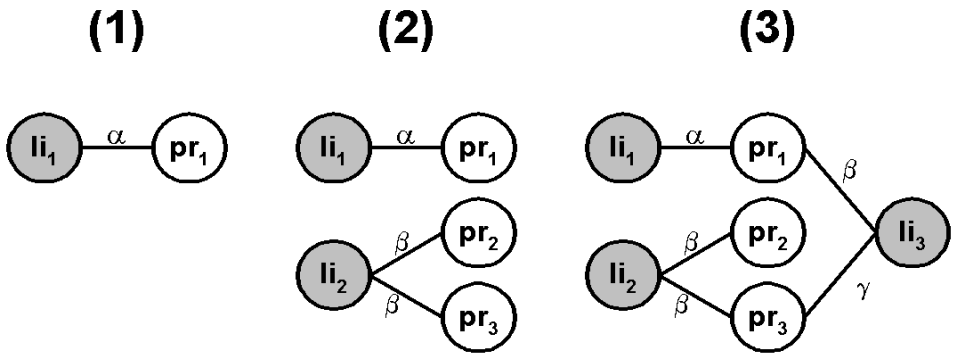

Figura 4.5: Inserção de referências geográficas no grafo de inferência geográfica.

criação do grafo de inferência em uma seqüência de três passos, cada um representando a inserção em G de uma das referências geográficas dadas como exemplo. G encontrava–se inicialmente vazio.

Agora que sabemos como criar um grafo de inferência geográfica, podemos descrever como utilizá–lo para derivar conclusões a respeito da posição aproximada entre dois vértices de G. Representando por $w(uv)$ o peso de uma aresta uv, temos a definição a seguir.

Definição 2 *Sejam v_i e v_{i+n} vértices em G tal que existe um caminho $v_i \rightarrow v_{i+1} \rightarrow \ldots \rightarrow v_{i+(n-1)} \rightarrow v_{i+n}$ de tamanho n entre eles. A distância entre v_i e v_{i+n} é dada pela seguinte relação de recorrência:*

- $d_1 = w(v_i v_{i+1})$
- $d_n = D(d_{n-1}, w(v_{i+(n-1)} v_{i+n}))$

A definição acima descreve como calcular a distância entre dois vértices em G, separados por um caminho de tamanho n. Para ilustrar, tomemos o grafo de inferência G do passo (3) da Figura 4.5 para calcular a a distância entre os vértices pr_1 e pr_3. O caminho entre esses vértices possui um tamanho $n = 2$. Portanto, a distância entre eles será dada por d_2. Veja:

- $d_1 = w(pr_1 li_3) = \beta$
- $d_2 = D(d_1, w(li_3 pr_3)) = D(\beta, \gamma)$

O que acabamos de fazer corresponde a determinar, de modo aproximado, a distância entre dois pontos de referência apenas com a informação da distância de cada um em relação ao local de interesse comum a ambos. Pela Definição 2, assim como por esse resultado, é possível perceber que o cálculo da distância entre os vértices depende

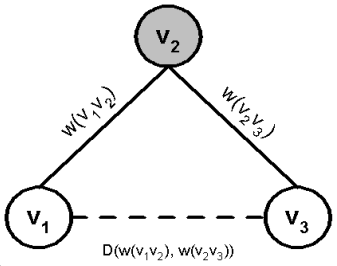

Figura 4.6: Representação gráfica da função D.

de D. Conforme mostra a Figura 4.6, D é uma função que calcula a distância entre dois vértices v_1 e v_3, conectados por uma caminho de tamanho $n = 2$ e adjacentes a um mesmo vértice v_2, com base nas distâncias $w(v_1v_2)$ e $w(v_2v_3)$. O valor retornado por D varia conforme a posição de v_1 e v_3 em relação a v_2. Como essas posições são desconhecidas, cabe à função D estipulá-las. A única restrição é que o valor retornado por D esteja entre $|w(v_1v_2) - w(v_2v_3)|$ e $w(v_1v_2) + w(v_2v_3)$, que correspondem respectivamente às distâncias mínima e máxima quando v_1 e v_3 estão posicionados ao longo de uma reta que passa por eles e por v_2 (Figura 4.7). É importante destacar que, como o cálculo das distâncias é feito de forma aproximada, quanto maior o tamanho n do caminho entre os vértices, menos precisa será a distância encontrada.

A seguir, definimos outra operação sobre o grafo de inferência geográfica, a função de seleção.

Definição 3 *Sejam v_i e v_{i+n} vértices em G tais como na Definição 2. A função $select(v_i, n, maxdist)$ retorna um conjunto de vértices v_j, $i + 1 \leq j \leq i + n$, tal que, para cada vértice desse conjunto, o valor da distância entre ele e v_i, dado pela relação de recorrência d, seja menor do que maxdist. Se v_j corresponde a um local de interesse, ele será descartado.*

Pela definição acima, vê-se que a função *select* retorna todos os pontos de referência v_j na vizinhança de v_i, cujo tamanho do caminho entre eles seja de no máximo n e cuja distância à v_i seja menor ou igual ao valor de *maxdist*. Os locais de interesse são descartados pois são locais desconhecidos, e, portanto, não possuem muita utilidade

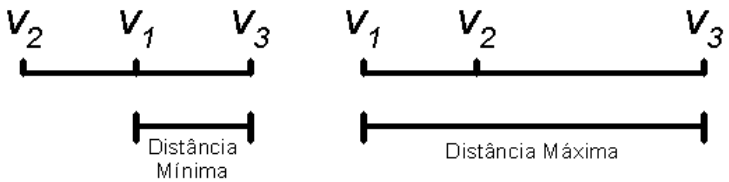

Figura 4.7: Distâncias euclidianas mínima e máxima entre os vértices v_1 e v_3.

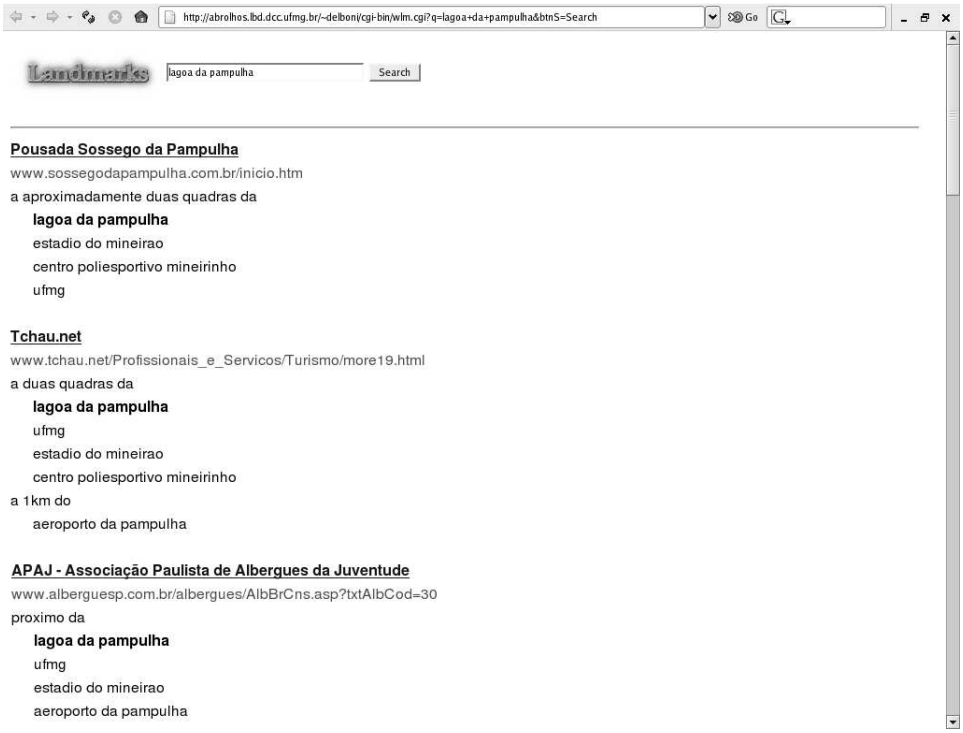

Figura 4.8: Exemplo de uma aplicação que utiliza a função *select*.

nesse momento. Com essa função é possível implementar uma aplicação de recuperação de informação geográfica onde v_i, n e *maxdist* constituem a consulta especificada pelo usuário. Opcionalmente, n e *maxdist* podem ser estipulados como parâmetros fixos da aplicação. A Figura 4.8 exibe a tela de resposta para a consulta 'lagoa da pampulha' em um protótipo de aplicação que utiliza a função *select*. Dado um ponto de referência, os locais próximos a ele retornados pela função *select* são exibidos juntamente com *links* para os documentos onde esses locais aparecem como pontos de referência, em expressões de posicionamento compostas ou em uma seqüência de expressões de posicionamento. Nessa aplicação, o parâmetro *maxdist* está sendo ignorado e, para a consulta mostrada, utilizamos $n = 2$.

Um outro exemplo de emprego dessa função seria na aplicação descrita na Seção 4.1, onde poderíamos utilizá–la para acrescentar novas expressões de posicionamento à consulta expandida. Se em uma consulta nessa aplicação o ponto de referência for definido como 'Lagoa da Pampulha', a função *select* pode ser utilizada para retornar locais próximos, tais como 'UFMG', 'Mineirão' e 'Aeroporto da Pampulha', que serão acrescentados à consulta expandida juntamente com as expressões de posicionamento contendo 'Lagoa da Pampulha'.

Como veremos a seguir, à medida que as informações provenientes de processos não cobertos neste trabalho tornam–se disponíveis, como a resolução de ambigüidade de nomes, a expressividade do grafo de inferência geográfica aumenta, o que permite

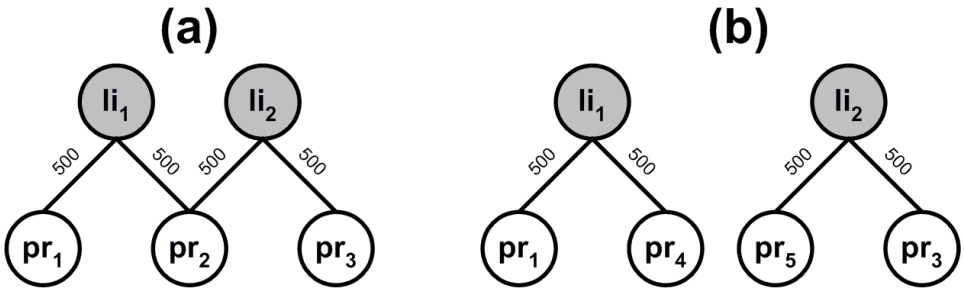

Figura 4.9: Grafo de Inferência do Exemplo 1

explorar melhor as expressões de posicionamento em aplicações de RIG. O nosso grafo de inferência, tal como apresentado até o momento, apresenta um problema relacionado aos locais genéricos (ex.: 'aeroporto', 'prefeitura') — eles são pontos de referência importantes, isto é, aparecem com freqüência e encontram–se associados a locais de interesse de vários lugares diferentes. Dessa forma, para casos onde $n > 3$, a função *select* pode recuperar dados incorretos, como mostra o Exemplo 1.

Exemplo 1 *Dado o grafo de inferência geográfica G da Figura 4.9a, os vértices pr_2 e pr_3 são retornados como resposta pela função select$(pr_1, 4, 2000)$. Para os valores das arestas fornecidos no exemplo, independente da função D utilizada, a resposta será sempre a mesma. Isso ocorre pois o valor máximo da distância euclidiana entre dois vértices de G foi atribuído ao parâmetro maxdist na função select. Entretanto, se pr_2 for um ponto de referência genérico, como 'aeroporto', ele pode referir-se ao 'Aeroporto da Pampulha' no contexto envolvendo o local de interesse li_1 e ao 'Aeroporto de Congonhas' no contexto de li_2. Nesse caso, os pontos de referência pr_1 e pr_3 estarão separados por uma distância de centenas de quilômetros, e não por uma distância máxima de 2000 metros. Esse fato também ocorre quando temos pontos de referência distintos, porém com o mesmo nome, pois são representados em G por um único vértice. Esse é o caso, por exemplo, da 'Praça Tiradentes' em Curitiba, Outro Preto, Belo Horizonte e em outros municípios brasileiros.*

A solução para ambos os casos descritos acima seria utilizar algum método de resolução de ambigüidade de nomes. No caso do Exemplo 1, se esse método tivesse sucesso em determinar quando pr_2 significa 'Aeroporto da Pampulha' e quando significa 'Aeroporto de Congonhas', pr_2 deveria ser dividido em dois vértices, pr_4 e pr_5. Dessa forma, G seria formado por dois grafos desconectados, como mostra a Figura 4.9b.

O emprego de métodos de resolução de ambigüidade permitem ainda que estratégias de expansão de consultas como as descritas na Seção 4.1 possam ser utilizadas em aplicações onde o ponto de referência é especificado como um lugar e o escopo geográfico

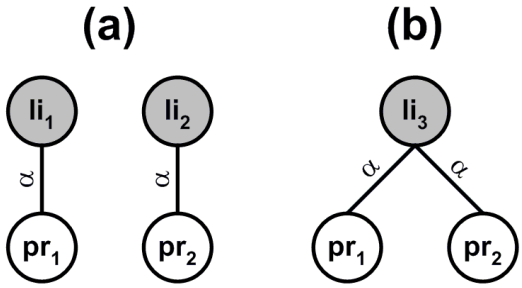

Figura 4.10: Grafo de Inferência do Exemplo 2

é a sua área. Por exemplo, uma consulta do tipo 'hotel em Belo Horizonte', poderia ser expandida para uma consulta booleana contendo várias expressões de posicionamento referentes a importantes pontos de referência em Belo Horizonte, como $q_{expandida}$ = (hotel AND ('próximo ao Minascentro' OR 'perto do Minascentro' OR ... OR 'próximo à UFMG' OR 'perto da UFMG' OR ... OR 'próximo ao BH Shopping' OR 'perto do BH Shopping' OR ...)).

Um outro problema refere–se à conectividade do grafo de inferência geográfica. Quanto mais conectado ele for, mais locais podem ser recuperados pela função *select*. Porém, como as expressões de posicionamento simples representam uma parcela bem mais significativa do que a de expressões de posicionamento compostas ou que aparecem em uma seqüência de expressões de posicionamento, a conectividade de G é baixa, apresentando um grande número de grafos K_2 desconectados, isto é, grafos completos com apenas dois vértices. Na coleção `WBR05`, por exemplo, 83,7% das expressões de posicionamento são do tipo simples. Para aumentar a conectividade do grafo, poderíamos utilizar um método para identificar os locais de interesse aos quais as expressões de posicionamento se referem. Dessa forma, os vértices de G que representam locais de interesse deixariam de ser tratados como locais "desconhecidos" e, ao invés de existir em G um vértice do tipo local de interesse para cada referência espacial (veja comentário após a Definição 1), existiria um vértice para cada local de interesse distinto. Veja o Exemplo 2 a seguir.

Exemplo 2 *Seja o G o grafo de inferência da Figura 4.10a, criado a partir das referências espaciais $\langle li_1, (\text{'perto de'}, pr_1)\rangle$ e $\langle li_2, (\text{'perto de'}, pr_2)\rangle$. Se utilizarmos um método para identificar os locais de interesse li_1 e li_2 e concluir que eles representam um mesmo local, G será reconfigurado: os vértices li_1 e li_2 se unirão para formar um único vértice, li_3, como na Figura 4.10b. Com isso, podemos dizer que li_3 está 'perto de' pr_1 e de pr_2 e que pr_1 e pr_2 encontram–se na vizinhança de li_3, estando portanto próximos um do outro, informação que não tínhamos na situação anterior (Figura 4.10a).*

Tabela 4.2: Nomes alternativos utilizados para designar um mesmo local.

Ocorrências	Nome
26	aeroporto internacional de salvador
25	aeroporto internacional luis eduardo magalhaes
19	aeroporto de salvador
14	aeroporto internacional luiz eduardo magalhaes
13	aeroporto internacional deputado luis eduardo magalhaes
12	aeroporto luis eduardo magalhaes
9	aeroporto internacional luiz eduardo de magalhaes de salvador
4	aeroporto luiz eduardo magalhaes
2	aeroporto deputado luis eduardo magalhaes
1	aeroporto luis magalhaes
1	aeroporto dois de julho
126	TOTAL

Uma outra maneira de aumentar a conectividade do grafo de inferência geográfica é identificando–se apelidos ou nomes alternativos atribuídos a um mesmo local. Amitay et al. (2004) denominam *aliasing* a existência de múltiplos nomes para um mesmo local. Os nomes na Tabela 4.2 foram obtidos na extração de expressões de posicionamento realizada na coleção `WBR05` e ilustram bem o problema de *aliasing* — todos os nomes referem–se ao aeroporto da cidade de Salvador/BA, cujo nome oficial é 'Aeroporto Internacional de Salvador – Deputado Luís Eduardo Magalhães'.

Finalmente, mostramos como coordenadas geográficas podem ser atribuídas a locais de um grafo de inferência geográfica a partir de um pequeno conjunto de locais georreferenciados. O procedimento descrito a seguir resume as etapas desse processo:

1. Atribua coordenadas aos pontos de referência — pode ser feito fisicamente, mediante georreferenciamento de um local, com auxílio de um GPS, por exemplo, ou encontrando e geocodificando o endereço do ponto de referência. A Figura 4.11 exibe a cobertura dos pontos de referência das expressões de posicionamento extraídas da coleção `WBR05`. Pelo gráfico, vemos que uma pequena parcela dos pontos de referência é suficiente para cobrir boa parte das expressões de posicionamento: 10% dos pontos de referência aparecem em cerca de 70% das expressões de posicionamento. Dessa forma, a geocodificação pode ser feita apenas para os pontos de referência mais importantes;

2. Utilize o grafo de inferência geográfica para determinar as coordenadas geográficas aproximadas dos locais de interesse — com as distâncias de um local de interesse a cada um dos seus pontos de referência e as coordenadas geográficas dos pontos

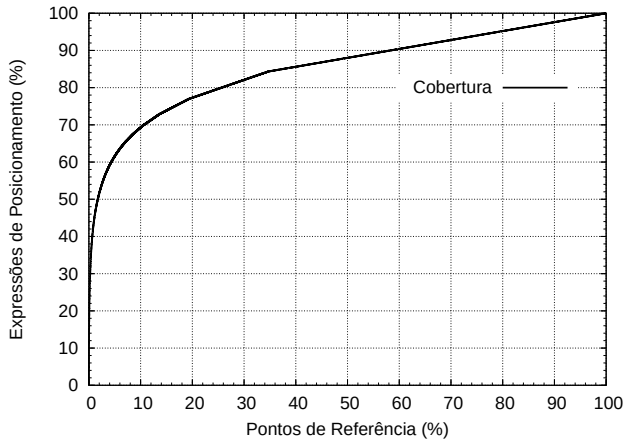

Figura 4.11: Cobertura das expressões de posicionamento pelos pontos de referência.

de referência, é possível determinar coordenadas aproximadas para a posição do local de interesse.

Podemos utilizar vários métodos para calcular de forma aproximada a posição do local de interesse. Por exemplo, se os valores de distância dados pela relação de recorrência d_n fossem exatos, poderíamos utilizar a interseção de circunferências para determinar a localização de um local de interesse, atribuindo–lhe coordenadas geográficas. Ao local de interesse seria atribuída uma coordenada precisa se ele estiver conectado em G a pelo menos três pontos de referência georreferenciados, conforme discutido a seguir.

Definição 4 *$C_{j,n} = (c_n, r_{j,n})$ é uma circunferência de centro c_n e raio $r_{j,n}$, onde c_n é dado pelas coordenadas (x_n, y_n) do n-ésimo ponto de referência pr_n adjacente a li_j e $r_{j,n} = d_1(li_j, pr_n)$.*

Definição 5 *$loc_m(li_j)$ é a função que determina as coordenadas geográficas aproximadas de li_j em função dos pontos de referência pr_n, $1 \leq n \leq m$, aos quais encontra–se conectado em G. De acordo com o grau m do vértice li_j, isto é, número de pontos de referência ao qual está conectado, temos:*

- *Para $m = 1$, $loc_1(li_j)$ retorna qualquer ponto sobre o perímetro de $C_{j,1}$, ou seja, qualquer coordenada (x, y) que satisfaça a equação $x^2 + y^2 = {r_{j,1}}^2$.*
- *Para $m = 2$, dois casos devem ser analisados: (1) Se $C_{j,1}$ e $C_{j,2}$ são circunferências tangentes, internas ou externas, $loc_2(li_j)$ retorna o ponto de interseção entre $C_{j,1}$ e $C_{j,2}$; e (2) Se $C_{j,1}$ e $C_{j,2}$ são circunferências secantes, $loc_2(li_j)$ retorna um dos pontos da interseção entre $C_{j,1}$ e $C_{j,2}$.*

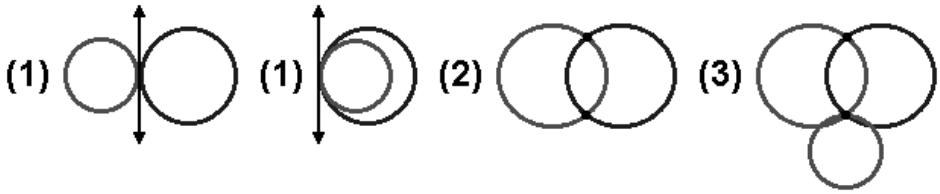

Figura 4.12: Interseções entre circunferências.

- *Para $m = 3$, $loc_3(li_j)$ retorna o ponto de interseção entre $C_{j,1}$, $C_{j,2}$ e $C_{j,3}$.*
- *Para $m > 3$, basta selecionar três pontos de referência quaisquer e executar o procedimento para $m = 3$.*

Os itens (1) e (2) da Figura 4.12 ilustram as situações possíveis onde $m = 2$, enquanto que o item (3) corresponde ao caso onde $m = 3$.

Como o valor $r_{j,n}$, fornecido por $d_1(li_j, pr_n)$ não é um valor preciso, e sim um valor médio calculado para uma pequena amostra de cada tipo de relação espacial, ele traz consigo uma margem de erro. Dessa forma, a abordagem descrita acima não seria adequada; é preciso encontrar métodos mais eficazes para determinar a localização de li_j de forma aproximada. A avaliação de tais métodos, porém, não é contemplada por este trabalho. Ao invés disso, descreve–se a seguir um exemplo prático com dados reais da cidade de Salvador, cujo objetivo é tão somente demonstrar a viabilidade em se atribuir coordenadas geográficas de forma aproximada a locais de interesse utilizando o grafo de inferência geográfica.

O *CarnaSite.com.br* é um *site* dedicado ao Carnaval de Salvador. Nele, é possível obter diversas informações sobre o evento, inclusive a respeito de onde se hospedar na cidade[4]. Utilizando a ferramenta DEByE (Laender et al., 2002), foi possível extrair 88 registros sobre locais de hospedagem, como hotéis, pousadas, apart–hotéis e albergues, como o exibido na Figura 4.13.

Em seguida, selecionamos todos os locais de hospedagem, num total de 55, cuja localização é descrita por expressões de posicionamento tendo o 'aeroporto', o 'centro' *e* a 'rodoviária' como pontos de referência. No exemplo da Figura 4.13 temos: "*a 5 minutos do centro*, *40 minutos do aeroporto* e *20 minutos da rodoviária*". A localização desses três pontos de referência foi geocodificada utilizando–se as coordenadas do centróide dos respectivos RMEs. Com esses dados, foi criado um grafo de inferência G, onde os locais de hospedagem são os locais de interesse e os pontos de referência são pr_1='aeroporto', pr_2='centro' e pr_3='rodoviária'. O valor de $d_1(li_j, pr_n)$ é dado pelo valor nominal X da relação espacial métrica 'a X minutos de' multiplicado pela

[4] `http://www.carnasite.com.br/carnaval/ondeficar.asp`

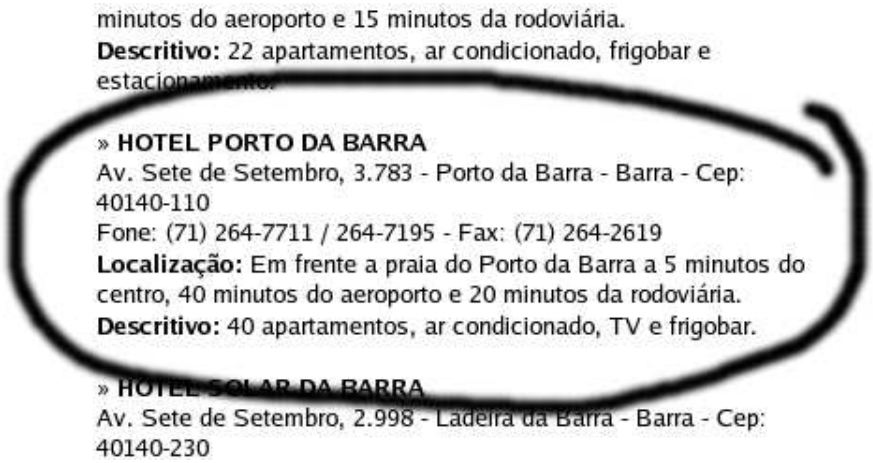

Figura 4.13: Registro com dados do *site* CarnaSite.

distância média correspondente a '1 minuto' encontrada na Seção 3.4, que é 530,90 metros.

Para encontrar a posição aproximada de li_j em relação a pr_1, pr_2 e pr_3, foi utilizado o seguinte procedimento:

1. Encontre os pontos de interseção das circunferências $C_{j,1}$, $C_{j,2}$ e $C_{j,3}$ duas a duas. Para cada par p de circunferências, $(C_{j,1}, C_{j,2})$, $(C_{j,1}, C_{j,3})$ e $(C_{j,2}, C_{j,3})$, temos as seguintes possibilidades:

   a) as circunferências são secantes — determine as coordenadas dos pontos de interseção.

   b) as circunferências são tangentes — determine a coordenada do ponto de interseção.

   c) as circunferências não se interceptam — essa situação é possível pois o peso da aresta li_jpr_n é um valor médio, e não exato. Nesse caso, como ilustra a Figura 4.14, não há pontos de interseção, e as circunferências podem ser (1) *externas* ou (2) *internas*. Entretanto, vamos interpolar um ponto de interseção da seguinte forma:

      i. Trace uma reta que passa pelos centros das circunferências;

      ii. Identifique as coordenadas dos pontos de interseção entre a reta e cada uma das circunferências;

      iii. Encontre as coordenadas do ponto médio do segmento de reta entre os pontos encontrados no item anterior. No caso das circunferências serem internas, há dois pontos de interseção da reta com a circunferência interior — o ponto mais próximo do ponto de interseção da reta com

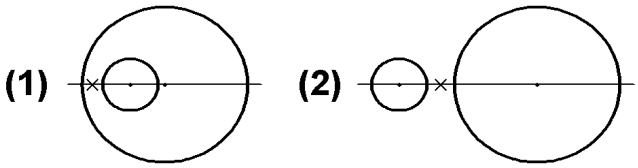

Figura 4.14: Interpolação do ponto de interseção entre as circunferências.

a outra circunferência deverá ser utilizado. O item (1) da Figura 4.14 ilustra essa situação.

2. A localização aproximada do local de interesse li_j será dada pela coordenada média dos pontos de interseção encontrados entre cada par p de circunferências $C_{j,n}$. Caso as circunferências de p possuam apenas um ponto de interseção (casos 1b e 1c), eles deverão receber peso 2 no cálculo da coordenada média.

Procedendo dessa forma, foi possível determinar uma coordenada geográfica aproximada para cada local de interesse li_j em G, como representado na Figura 4.15. Nela, três circunferências, ora tangentes ora secantes entre si, determinam com exatidão a localização de um hotel fictício na cidade de Salvador. Nessa mesma figura, os hotéis utilizados no experimento estão retratados pelos ícones distribuídos ao longo do litoral.

Medindo–se a distância entre a coordenada aproximada e a coordenada real, é possível avaliar a qualidade do método em termos de precisão. No caso desse experimento,

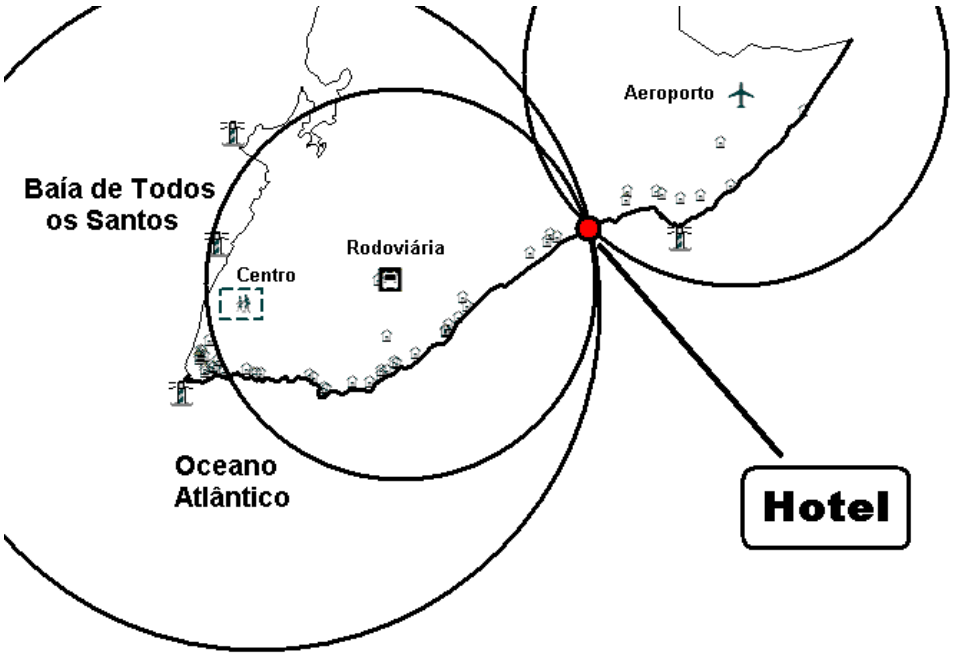

Figura 4.15: As interseções das circunferências obtidas a partir do grafo de inferência determinam as coordenadas de um local de interesse em Salvador.

a distância média observada foi de 3,6 km. A princípio, poderia–se duvidar da eficácia do método, pois esse valor pode ser considerado alto. Entretanto, vamos fazer uma análise mais cuidadosa. Os pontos de referência utilizados encontram–se distantes uns dos outros e, em muitos casos, afastados também dos locais de hospedagem: a distância entre esses locais e os pontos de referência varia de 291 metros a 23,6 km, com uma média de 10,6 km. Dessa forma, a área ao redor da coordenada aproximada é quase 9 vezes menor do que a área média ao redor de um ponto de referência, ou seja, a coordenada aproximada, mesmo com um erro médio de 3,6 km em relação à coordenada real, estabelece uma área bem mais precisa do que aquela determinada, na média, por uma única expressão de posicionamento. Isso nos faz acreditar que o cálculo de coordenadas geográficas aproximadas utilizando o método proposto e o grafo de inferência é valido e que, se os pontos de referência utilizados fossem mais próximos dos locais de interesse, o resultado obtido poderia ter sido bem melhor. Não obstante, outros métodos devem ser pesquisados, o que certamente irá contribuir para melhorar a qualidade das coordenadas calculadas.

Com esse exemplo, demonstramos como obter as coordenadas geográficas aproximadas de locais de interesse a partir das coordenadas de pontos de referência adjacentes a eles em um grafo de inferência. Entretanto, essa abordagem pode ser utilizada para calcular as coordenadas geográficas aproximadas de qualquer vértice de um grafo de inferência, desde que exista um caminho entre esse vértice e vértices associados à coordenadas geográficas. Nesse caso, é importante ressaltar que quanto maior o tamanho do caminho entre o vértice georreferenciado e o vértice para o qual deseja–se calcular as coordenadas, mais imprecisa serão as coordenadas obtidas.

Capítulo 5

Conclusões e Trabalhos Futuros

Neste trabalho foi apresentado um estudo que buscou caracterizar as *expressões de posicionamento*, uma fonte de evidência de contexto geográfico encontrada na *Web*, porém pouco explorada pela literatura. Através de um estudo de caso da cidade de Belo Horizonte, foi possível determinar quais os tipos de local mais usados como ponto de referência e como local de interesse, as relações espaciais mais importantes e que, em quase 82% das ocorrências, as expressões de posicionamento são utilizadas como forma de descrever a posição no espaço de um local físico.

Um extrator capaz de identificar expressões de posicionamento em documentos textuais também foi desenvolvido. Devido à falta de informações sobre locais intra–urbanos nos atuais *gazetteers*, decidimos por uma abordagem que dispensa a utilização desses repositórios de nomes, embora não impeça seu uso caso estejam disponíveis. Os nomes de locais são reconhecidos por regras sintáticas dentro do contexto fornecido por uma relação espacial, o que torna desnecessária a utilização de métodos de resolução de ambigüidades do tipo *geo/não–geo*. Como efeito colateral positivo, um repositório de nomes de locais é gerado como resultado da análise dos documentos pelo extrator implementado para esse fim. Cerca de 90% dos nomes desse repositório correspondem a nomes de locais (Seção 3.3).

As expressões de posicionamento em aplicações de busca local demonstraram possuir um potencial para desempenhar um papel tão importante quanto o das fontes de contexto geográfico consideradas "tradicionais". Uma vantagem da expressão de posicionamento, como demonstrado no Capítulo 4, é que ela pode ser empregada em aplicações de RIG sem que a localização determinada pelo par ⟨*relação espacial, ponto de referência*⟩ precise ser convertida em coordenadas geográficas, isto é, a utilização de bancos de dados geográficos é dispensável. Nos países onde fontes de informação desse tipo estão em estágio inicial de criação, incompletas ou desatualizadas, a utilização de expressões de posicionamento pode trazer grande benefício. A aplicação de busca

local baseada em expansão de consultas apresentada na Seção 4.1, onde várias expressões de posicionamento compatíveis com o escopo geográfico da consulta são utilizadas como sinônimos, ilustra bem esse ponto: as consultas expandidas apresentaram uma precisão média cerca de 60% superior à das consultas originais. Essa aplicação mostra ainda que os pontos de referência podem ser utilizados como centro de busca na especificação do escopo geográfico das consultas, algo incompreensivelmente raro de se ver em aplicações desse tipo e que, assim como a utilização de coordenadas geográficas, a resolução de ambigüidade de nomes não é estritamente necessária para se construir uma aplicação de busca local. Nessa aplicação, confiamos ao algoritmo de *ranking* da máquina de busca utilizada para processar as consultas a tarefa de resolver possíveis ambigüidades no nome do local informado como ponto de referência.

Obviamente, como discutido na Seção 4.2, que introduziu o conceito de um grafo de inferência geográfica, quanto mais informações tivermos a respeito das expressões de posicionamento, poderemos utilizá–las de forma mais precisa e em mais aplicações. Informações provenientes de métodos de resolução de ambigüidade, geocodificação e identificação de locais de interesse contribuiriam muito nesse aspecto; demonstramos, porém, que a ausência delas pode ser suprida parcialmente por operações definidas sobre o grafo de inferência geográfica. Com elas é possível, por exemplo, encontrar locais próximos, sem que o relacionamento entre eles esteja explícito em alguma referência geográfica e até mesmo calcular de forma aproximada coordenadas geográficas a partir de locais cuja posição encontra–se geocodificada.

Futuros desenvolvimentos relacionados ao trabalho apresentado incluem:

1. Analisar outras relações espaciais — neste trabalho, apenas 30 relações espaciais foram utilizadas. Expressões de posicionamento como 'Rua Curitiba esquina com Av. Bias Fortes' ou 'Rua Alagoas, entre Antônio de Albuquerque e Fernandes Tourinho', onde aparecem relações espaciais de interseção, podem fornecer uma localização bastante precisa, principalmente se uma base de endereços georreferenciados estiver disponível. Além disso, as relações espaciais utilizadas foram escolhidas entre as que julgamos ser as principais representantes das categorias *fuzzy*, métrica, direcional e topológica, mas certamente podem existir outras não consideradas que também são importantes. Talvez seja possível encontrar novas relações espaciais da seguinte forma: (1) utilizar os nomes de locais presentes nas expressões de posicionamento identificadas por nosso extrator para localizar, nos documentos, os trechos onde eles ocorrem; (2) determinar de forma automática os padrões mais freqüentes que ocorrem antes dos nomes de locais; e (3) fazer uma inspeção nos padrões encontrados para tentar identificar relações espaciais entre eles.

2. Realizar uma avaliação mais completa das expressões de posicionamento encontradas por nosso extrator — neste trabalho, realizamos uma análise global das expressões de posicionamento extraídas, classificando uma amostra aleatória na qual verificamos que 89,6 ± 4,0 % estão corretas. É preciso (1) classificar uma amostra maior para diminuir a margem de erro e obter um valor mais preciso; e (2) avaliar individualmente cada relação espacial, identificando a contribuição de cada uma para a extração como um todo em termos de extrações corretas e incorretas. Com isso, podemos tentar melhorar a qualidade da extração, seja eliminando do algoritmo relações espaciais inadequadas, isto é, com baixo percentual de extrações corretas, seja identificando palavras freqüentemente confundidas com pontos de referência e desconsiderando as expressões de posicionamento onde elas aparecem.

3. Resolver ambigüidades — como existem poucos dados de locais intra–urbanos do Brasil disponíveis em *gazetteers*, é preciso desenvolver um método para resolver ambigüidades de nomes de locais que não dependa desses repositórios. Seguindo a premissa de que os pontos de referência encontrados por nosso extrator referem–se a locais, não necessitando, portanto, de tratamento para ambigüidades do tipo *geo/não–geo*, podemos utilizar a seguinte estratégia: (1) Caso exista um *gazetteer* disponível, utilizá–lo para resolver nomes únicos de imediato; (2) Alguns pontos de referência já possuem explícito em seus nomes o lugar ao qual pertencem, como 'Santa Casa de Salvador' e 'centro de Porto Alegre'. Se não existirem duas ou mais cidades com o mesmo nome, a ambigüidade já estará resolvida; e (3) outras evidências geográficas existentes no documento, especialmente próximas ao trecho onde o ponto de referência foi encontrado, como CEPs, endereços, números de telefone com código DDD, além de nomes de locais já "resolvidos" podem ser utilizadas para inferir o lugar relacionado ao ponto de referência, inclusive com o auxílio de um grafo de inferência.

4. Criar um *gazetteer* — os nomes de locais "resolvidos", conforme discutido no item anterior, podem ser utilizados como um ponto de partida para a criação de um *gazetteer* onde, segundo Hill (2000), três atributos são essenciais para qualquer registro: nome, localização geográfica (*footprint*) e tipo. O nome do local é obtido trivialmente; a localização pode ser inicialmente atribuída a uma posição aleatória dentro da área do município ao qual pertence; e o tipo de grande parte dos pontos de referência é determinado por palavras como 'parque', 'igreja', 'praça', 'escola', 'shopping', 'hotel', 'hospital', etc. Os itens descritos a seguir podem ser utilizados para ampliar e melhorar a qualidade das informações do *gazetteer*:

- A extração de dados semi–estruturados de locais intra–urbanos existentes em guias de cidade e catálogos de serviço *on–line* pode ser utilizada para alimentar o *gazetteer*, como mostra Souza et al. (2005).
- Identificar locais de interesse — determinar os locais de interesse aos quais se referem as expressões de posicionamento encontradas em documentos da *Web* não é tarefa fácil. Além de poderem estar localizados em qualquer lugar ao longo do texto, geralmente outros nomes de locais também são mencionados no documento. Entretanto, procurar por tipos de local propícios a figurarem como locais de interesse em uma referência geográfica, tais como 'hotéis' ou 'restaurantes', parece ser um bom começo.
- Tratar o problema de *aliasing* — um mesmo local pode possuir vários nomes. O emprego de um método para identificar conjuntos de nomes possivelmente relacionados a um mesmo local certamente contribui para aumentar a qualidade das informações de um *gazetteer*. Estratégias semelhantes à apresentada por Oliveira et al. (2005) podem ser utilizadas se adaptadas para o contexto dos nomes de locais.
- Geocodificar pontos de referência e locais de interesse — os locais provenientes de registros de guias e catálogos *on–line* podem ter o seu endereço geocodificado. Além disso, como visto na Seção 4.2 o grafo de inferência geográfica pode ser utilizado para calcular de forma aproximada a coordenada geográfica de um local a partir das coordenadas de outros locais.

  Quanto mais locais geocodificados, maior a probabilidade de termos referências geográficas completamente geocodificadas, isto é, tanto o local de interesse quanto o ponto de referência da expressão de posicionamento associados a coordenadas geográficas. Com isso, os valores de distância atribuídos às relações espaciais podem ser continuamente ajustados. Pode–se, inclusive, calcular as distâncias para um região específica, um estado por exemplo, de modo a obter valores adequados para a realidade desse lugar, determinada por fatores como tamanho do território, densidade demográfica e questões culturais.

5. Formalizar o arcabouço de inferência geográfica — o grafo de inferência geográfica apresentado neste trabalho, juntamente com suas operações, pode ser transformado em um arcabouço de inferência geográfica. Para isso, é necessário uma definição mais rigorosa das operações e estruturas apresentadas, a introdução de novas operações e a proposição dos algoritmos necessários para a sua implementação.

Assim como para o *gazetteer*, praticamente todos os itens descritos acima trazem algum tipo de benefício para o grafo de inferência geográfica, que ganha em expressividade — outros tipos de inferências e operações podem ser realizadas — e em qualidade — os resultados obtidos são mais precisos. Isso sugere uma forte interação entre os dois, fato que deve ser melhor explorado.

Por fim, acreditamos ser preciso desenvolver mecanismos propícios para a integração da informação geográfica oriunda das expressões de posicionamento em ferramentas de busca local, assim como identificar outras aplicações de RIG que poderiam se beneficiar do uso dessa fonte de contexto geográfico, como, por exemplo, as aplicações de roteamento. Atualmente, existem disponíveis na Internet aplicações que, dadas duas coordenadas geográficas, especificadas por meio de algum localizador, como um endereço, fornece uma rota descritiva entre dois pontos, podendo–se, geralmente, escolher entre a menor rota ou a rota mais rápida. A descrição da rota é feita sob a forma de um conjunto de instruções em linguagem natural envolvendo distâncias, direções e nomes de logradouros como no trecho '(...) siga por *200 m* e vire *à direita* na *Rua XYZ*'. Essa descrição poderia mencionar *pontos de referência*, o que facilitaria a compreensão da rota, como verificado por Tom e Denis (2003).

Apêndice A

Relações Espaciais como Expressões Regulares

A.1 Fuzzy

A.1.1 "próximo ao"

```
‘\W(proxim[oa] d[oa]) ’
```

A.1.2 "perto de"

```
‘\W(perto d[oa]) ’
```

A.1.3 "depois de"

```
‘\W(depois d[oa]) ’
```

A.1.4 "antes de"

```
‘\W(antes d[oa]) ’
```

A.1.5 "nas proximidades de"

```
‘\W(nas proximidades d[oa]) ’
```

A.1.6 "abaixo de"

```
‘\W(abaixo d[oa]) ’
```

A.1.7 "pertinho de"

```
‘\W(pertinho d[oa]) ’
```

A.1.8 "acima de"

```
‘\W(acima d[oa]) ’
```

A.1.9 "na vizinhança de"

```
‘\W(nas? vizinhancas? d[oa]) ’
```

A.2 Direcionais

A.2.1 "em frente ao"

```
‘\W(em frente ao?) ’
```

A.2.2 "ao lado de"

```
‘\W(ao lado d[oa]) ’
```

A.2.3 "atrás de"

```
‘\W(atras d[oa]) ’
```

A.2.4 "defronte ao"

```
‘\W(defronte ao?) ’
```

A.3 Métricas

A.3.1 "a ? km de"

```
‘\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| a
prox(?:imadamente|.?)| uns| cerca de| (?:pouco )?m(?:ais|enos) de| quas
e| exatos| mais ou menos| [+][-])? (?:[[:alpha:]]+ |[[:digit:]][[:digit
:],.]* ?)km d(?:e distancia d[oa]|[oa])) ’
```

A.3.2 "a ? minutos de"

```
'\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| a
prox(?:imadamente|.?)| uns| cerca de| (?:pouco )?m(?:ais|enos) de| quas
e| exatos| mais ou menos| [+][-])? (?:[[:alpha:]]+|[[:digit:]]+) minuto
s? d(?:e (?:carro|caminhada|trem|bonde|metro|onibus|barco) d[oa]|[oa])) '
```

A.3.3 "a ? quilômetros de"

```
'\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| a
prox(?:imadamente|.?)| uns| cerca de| (?:pouco )?m(?:ais|enos) de| quas
e| exatos| mais ou menos| [+][-])? (?:[[:alpha:]]+|[[:digit:]][[:digit:
],.]*) quilometros? d(?:e distancia d[oa]|[oa])) '
```

A.3.4 "a ? metros de"

```
'\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| a
prox(?:imadamente|.?)| uns| cerca de| (?:pouco )?m(?:ais|enos) de| quas
e| exatos| mais ou menos| [+][-])? (?:[[:alpha:]]+|[[:digit:]][[:digit:
],.]*) metros? d(?:[oa]|e distancia d[oa])) '
```

A.3.5 "a ? quadras de"

```
'\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| a
prox(?:imadamente|.?)| umas| cerca de| (?:pouco )?m(?:ais|enos) de| qua
se| exatos| mais ou menos| [+][-])? (?:[[:alpha:]]+|[[:digit:]]+) quadr
as? d[oa]) '
```

A.3.6 "a ? m de"

```
'\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| a
prox(?:imadamente|.?)| uns| cerca de| (?:pouco )?m(?:ais|enos) de| quas
e| exatos| mais ou menos| [+][-])? (?:[[:alpha:]]+ |[[:digit:]][[:digit
:],.]* ?)m d(?:[oa]|e distancia d[oa])) '
```

A.3.7 "a ? min de"

```
'\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| a
prox(?:imadamente|.?)| uns| cerca de| (?:pouco )?m(?:ais|enos) de| quas
e| exatos| mais ou menos| [+][-])? (?:[[:alpha:]]+ |[[:digit:]]+ ?)min[
```

```
.]? d(?:e (?:carro|caminhada|trem|bonde|metro|onibus|barco) d[oae]|[oae
])) '
```

A.3.8 "a ? quarteirões de"

```
'\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| a
prox(?:imadamente|.?)| uns| cerca de| (?:pouco )?m(?:ais|enos) de| quas
e| exatos| mais ou menos| [+][-])? (?:[[:alpha:]]+|[[:digit:]]+) quarte
ir(?:ao|oes) d[oa]) '
```

A.3.9 "a ? blocos de"

```
'\W((?:a(?: uma distancia de)?|dista(?:nte|ndo)?)(?: apenas| somente| a
prox(?:imadamente|.?)| uns| cerca de| (?:pouco )?m(?:ais|enos) de| quas
e| exatos| mais ou menos| [+][-])? (?:[[:alpha:]]+|[[:digit:]]+) blocos
? d[oa]) '
```

A.4 Topológicas

A.4.1 "dentro de"

```
'\W(dentro (?:d[oa]|das dependencias d[oa])) '
```

A.4.2 "no coração de"

```
'\W(no coracao d[oa]) '
```

A.4.3 "no ? andar de"

```
'\W(no (?:[[:alnum:]]+ andar|andar [[:alnum:]]+) d[oa]) '
```

A.4.4 "em cima de"

```
'\W(em cima d[oa]) '
```

A.4.5 "no ? piso de"

```
'\W(no (?:[[:alnum:]]+ piso|piso [[:alnum:]]+) d[oa]) '
```

A.4.6 "embaixo de"

```
‘\W(embaixo d[oa]) ’
```

A.4.7 "na praça de alimentação de"

```
‘\W(na praca de alimentacao d[oa]) ’
```

A.4.8 "no ? nível de"

```
‘\W(no (?:[[:alnum:]]+ nivel|nivel [[:alnum:]]+) d[oa]) ’
```

Apêndice B

Ocorrência das Relações Espaciais

B.1 Relações Espaciais nas Expressões de Posicionamento Válidas e Inválidas

A coluna 'TOTAL' refere–se às relações espaciais presentes nas expressões de posicionamento extraídas da coleção de documentos utilizada no estudo de caso apresentado na Seção 3.2 e as colunas 'INVÁLIDAS' e 'VÁLIDAS', respectivamente, referem–se às relações espaciais das expressões de posicionamento consideradas inválidas ou válidas durante o processo de classificação. As colunas 'DM' exibem a distância média, em palavras, entre a relação espacial e o ponto de referência nas expressões de posicionamento.

As porcentagens da coluna 'TOTAL' devem ser interpretadas na vertical — tanto as ocorrências das categorias quanto das relações espaciais individuais somam 100%. As colunas 'INVÁLIDAS' e 'VÁLIDAS', entretanto, devem ser interpretadas na horizontal.

RELAÇÃO ESPACIAL	TOTAL		INVÁLIDAS		DM	VÁLIDAS		DM
TODAS	4889	100.00%	517	10.57%	2.93	4372	89.43%	0.17
FUZZY	1759	35.98%	271	15.41%	3.18	1488	84.59%	0.28
próximo ao	992	20.29%	14	1.41%	1.79	978	98.59%	0.32
perto de	367	7.51%	31	8.45%	2.26	336	91.55%	0.23
depois de	135	2.76%	114	84.44%	3.29	21	15.56%	0.00
antes de	108	2.21%	90	83.33%	3.46	18	16.67%	0.00
nas proximidades de	91	1.86%	1	1.10%	4.00	90	98.90%	0.14

abaixo de	24	0.49%	6	25.00%	3.83	18	75.00%	0.14
pertinho de	21	0.43%	1	4.76%	0.00	20	95.24%	0.43
acima de	21	0.43%	14	66.67%	3.93	7	33.33%	0.00
na vizinhança de	0	0.00%	0	0.00%	0.00	0	0.00%	0.00
DIRECIONAIS	1285	26.28%	39	3.04%	2.85	1246	96.96%	0.07
em frente ao	743	15.20%	7	0.94%	2.14	736	99.06%	0.09
ao lado de	399	8.16%	17	4.26%	2.47	382	95.74%	0.04
atrás de	111	2.27%	15	13.51%	3.60	96	86.49%	0.00
defronte ao	32	0.65%	0	0.00%	0.00	32	100.00%	0.00
MÉTRICA	1263	25.83%	1	0.08%	3.00	1262	99.92%	0.08
a ? km de	648	13.25%	0	0.00%	0.00	648	100.00%	0.00
a ? minutos de	229	4.68%	0	0.00%	0.00	229	100.00%	0.14
a ? quilômetros de	216	4.42%	0	0.00%	0.00	216	100.00%	0.00
a ? metros de	73	1.49%	1	1.37%	3.00	72	98.63%	0.00
a ? quadras de	41	0.84%	0	0.00%	0.00	41	100.00%	0.24
a ? m de	32	0.65%	0	0.00%	0.00	32	100.00%	0.00
a ? min de	12	0.25%	0	0.00%	0.00	12	100.00%	0.00
a ? quarteirões de	11	0.22%	0	0.00%	0.00	11	100.00%	0.00
a ? blocos de	1	0.02%	0	0.00%	0.00	1	100.00%	0.00
TOPOLÓGICA	582	11.90%	206	35.40%	2.61	376	64.60%	0.13
dentro de	399	8.16%	183	45.86%	2.60	216	54.14%	0.17
no coração de	66	1.35%	6	9.09%	2.67	60	90.91%	0.00
no ? andar de	59	1.21%	0	0.00%	0.00	59	100.00%	0.25
em cima de	20	0.41%	13	65.00%	3.23	7	35.00%	0.00
no ? piso de	15	0.31%	0	0.00%	0.00	15	100.00%	0.00
embaixo de	13	0.27%	4	30.77%	1.25	9	69.23%	0.00

na praça de alimentação de	7	0.14%	0	0.00%	0.00	7 100.00%	0.00
no ? nível de	3	0.06%	0	0.00%	0.00	3 100.00%	0.00

B.2 Relações Espaciais nas Expressões de Posicionamento Válidas

A coluna 'TOTAL' refere–se às relações espaciais presentes nas expressões de posicionamento válidas, a coluna 'VAL BH' às relações espaciais das expressões de posicionamento válidas cujos pontos de referência localizam–se na cidade de Belo Horizonte e a coluna 'VAL OUTRAS' às relações espaciais das expressões de posicionamento válidas cujos pontos de referência encontram–se em outras cidades.

RELAÇÃO ESPACIAL	TOTAL	VAL BH		VAL OUTRAS	
TODAS	4372	909	20.79%	3463	79.21%
FUZZY	1488	391	26.28%	1097	73.72%
próximo ao	978	272	27.81%	706	72.19%
perto de	336	74	22.02%	262	77.98%
depois de	21	4	19.05%	17	80.95%
antes de	18	5	27.78%	13	72.22%
nas proximidades de	90	21	23.33%	69	76.67%
abaixo de	18	7	38.89%	11	61.11%
pertinho de	20	7	35.00%	13	65.00%
acima de	7	1	14.29%	6	85.71%
na vizinhança de	0	0	0.00%	0	0.00%
DIRECIONAIS	1246	273	21.91%	973	78.09%
em frente ao	736	168	22.83%	568	77.17%
ao lado de	382	80	20.94%	302	79.06%
atrás de	96	22	22.92%	74	77.08%

defronte ao	32	3	9.38%	29	90.62%
MÉTRICA	1262	92	7.29%	1170	92.71%
a ? km de	648	23	3.55%	625	96.45%
a ? minutos de	229	22	9.61%	207	90.39%
a ? quilômetros de	216	12	5.56%	204	94.44%
a ? metros de	72	5	6.94%	67	93.06%
a ? quadras de	41	17	41.46%	24	58.54%
a ? m de	32	4	12.50%	28	87.50%
a ? min de	12	0	0.00%	12	100.00%
a ? quarteirões de	11	9	81.82%	2	18.18%
a ? blocos de	1	0	0.00%	1	100.00%
TOPOLÓGICA	376	153	40.69%	223	59.31%
dentro de	216	84	38.89%	132	61.11%
no coração de	60	24	40.00%	36	60.00%
no ? andar de	59	24	40.68%	35	59.32%
em cima de	7	2	28.57%	5	71.43%
no ? piso de	15	9	60.00%	6	40.00%
embaixo de	9	2	22.22%	7	77.78%
na praça de alimentação de	7	6	85.71%	1	14.29%
no ? nível de	3	2	66.67%	1	33.33%

Referências Bibliográficas

Aho, A. V. e Corasick, M. J. (1975). Efficient String Matching: an Aid to Bibliographic Search. *Communications of the ACM*, 18(6):333–340.

Amitay, E.; Har'EI, N.; Sivan, R. e Soffer, A. (2004). Web–a–Where: Geotagging Web Content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 273–280, Sheffield, UK.

Borges, K. A. V.; Laender, A. H. F.; Medeiros, C. B.; da Silva, A. S. e Davis Jr., C. A. (2003). The Web as a Data Source for Spatial Databases. In *Proceedings of the V Brazilian Symposium on GeoInformatics*, Campos do Jordão, Brazil. Em CD–ROM. Disponível em `http://www.geoinfo.info`. Acessado em agosto de 2005.

Borges, K. A. V. e Sahay, S. (2000). GIS for the Public Sector: Experiences from the City of Belo Horizonte. *Information Infrastructure and Policy*, 6(3):139–155.

Buyukkokten, O.; Cho, J.; Garcia–Molina, H.; Gravano, L. e Shivakumar, N. (1999). Exploiting Geographical Location Information of Web Pages. In *Proceedings of the Workshop on Web Databases*, pp. 91–96, Philadelphia, USA. Held in conjunction with ACM SIGMOD.

Cunningham, H.; Maynard, D.; Bontcheva, K. e Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia, USA.

Delboni, T. M.; Borges, K. A. V. e Laender, A. H. F. (2005). Geographic Web Search based on Positioning Expressions. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, pp. 61–64, Bremen, Germany. Held in conjunction with ACM CIKM.

Ding, J.; Gravano, L. e Shivakumar, N. (2000). Computing Geographical Scopes of Web Resources. In *Proceedings of the 26th International Conference on Very Large Data Base*, pp. 545–556, Cairo, Egypt.

Egenhofer, M. J. e Franzosa, R. D. (1991). Point–set Topological Spatial Relations. *International Journal of Geographical Information Systems*, 2(5):161–174.

Egenhofer, M. J.; Mark, D. M. e Herring, J. (1994). The 9-Intersection: Formalism and Its Use for Natural–Language Spatial Predicates. Technical Report 94–1, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA, USA.

Egenhofer, M. J. e Shariff, R. B. M. (1998). Metric Details for Natural–Language Spatial Relations. *ACM Transactions on Information Systems*, 14(5):295–321.

Freeman, J. (1975). The Modelling of Spatial Relations. *Computer Graphics and Image Processing*, 4(2):156–171.

Guting, R. H. (1994). An Introduction to Spatial Database Systems. *The VLDB Journal*, 3(4):357–400.

Heinzle, F.; Kopczynski, M. e Sester, M. (2003). Spatial Data Interpretation for the Intelligent Access to Spatial Information in the Internet. In *Proceedings of the 21th International Cartographic Conference*, Durban, Africa.

Hill, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 280–290, London, UK. Springer–Verlag.

Himmelstein, M. (2005). Local Search: The Internet is the Yellow Pages. *IEEE Computer*, 38(2):26–35.

IEEE (2001). *IEEE Std 1003.1-2001 Standard for Information Technology — Portable Operating System Interface (POSIX) Base Definitions, Issue 6.* IEEE Computer Society, 345 E. 47th St, New York, NY 10017, USA.

Kohler, J. (2003). Analysing Search Engine Queries for the Use of Geographic Terms. Master's thesis, University of Sheffield, Sheffield, UK.

Kuhn, W.; Worboys, M. F. e Timpf, S., editores (2003). *Spatial Information Theory. Foundations of Geographic Information Science, International Conference, COSIT 2003, Ittingen, Switzerland, September 24-28, 2003, Proceedings*, volume 2825 of *Lecture Notes in Computer Science*. Springer.

Laender, A. H. F.; Ribeiro-Neto, B. e da Silva, A. S. (2002). DEByE — Data Extraction by Example. *Data and Knowledge Engineering*, 40(2):121–154.

Larson, R. R. (1996). Geographic Information Retrieval and Spatial Browsing. In Smith, L. e Gluck, M., editores, *GIS and Libraries: Patrons, Maps and Spatial Information*, pp. 81–124. University of Illinois, Urbana–Champaign, IL.

McCurley, K. S. (2001). Geospatial Mapping and Navigation of the Web. In *Proceedings of the 10th International World Wide Web Conference*, pp. 221–229, Hong Kong, China.

Montello, D.; Goodchild, M.; Gottsegen, J. e Fohl, P. (2003). Where's Downtown? Behavioral Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition and Computation*, 3(2,3):185–204.

Moratz, R. e Wallgrün, J. O. (2003). Spatial Reasoning about Relative Orientation and Distance for Robot Exploration. In Kuhn et al. (2003), pp. 61–74.

Oliveira, J. W.; Laender, A. H. F. e Gonçalves, M. A. (2005). Remoção de Ambigüidades na Identificação de Autoria de Objetos Bibliográficos. In *Anais do XX Simpósio Brasileiro de Banco de Dados*, Uberlândia, MG.

Papadias, D. e Sellis, T. (1994). Qualitative Representation of Spatial Knowledge in Two–Dimensional Space. *The VLDB Journal*, 3(4):479–516.

Pasca, M. (2004). Acquisition of Categorized Named Entities for Web Search. In *Proceedings of the 13th ACM Conference on Information and Knowledge Management*, pp. 137–145, Washington, D.C., USA.

Pullar, D. e Egenhofer, M. J. (1988). Towards the Defaction and Use of Topological Relations Among Spatial Objects. In *Proceedings of the 3rd International Symposium on Spatial Data Handling*, pp. 225–242, Columbus, Ohio.

Rauch, E.; Dukatin, M. e Baker, K. (2003). A Confidence–Based Framework for Disambiguating Geographic Terms. In *HLT–NAACL 2003 Workshop on Analysis of geographic References*, pp. 50–54, Edmonton, Canada.

Rodríguez-Tastets, M. A. (2002). A Spatial Dimension for Searching the World Wide Web. In *Proceedings of the Hybrid Intelligent Systems 2002*, pp. 583–592, Santiago, Chile.

Sanderson, M. e Kohler, J. (2004). Analyzing Geographic Queries. In *Proceedings of the 2004 Workshop on Geographic Information Retrieval*, Sheffield, UK. Held in conjunction with ACM SIGIR.

Silva, M. J.; Martins, B.; Chaves, M.; Cardoso, N. e Afonso, A. P. (2004). Adding Geographic Scopes to Web Resources. In *Proceedings of the 2004 Workshop on Geographic Information Retrieval*, Sheffield, UK. Held in conjunction with ACM SIGIR.

Souza, L. A.; Davis Jr., C. A.; Borges, K. A. V.; Delboni, T. M. e Laender, A. H. F. (2005). The Role of Gazetteers in Geographic Knowledge Discovery on the Web. In *Proceedings of the 3rd Latin American Web Congress*, pp. 157–165, Buenos Aires, Argentina.

Tom, A. e Denis, M. (2003). Referring to Landmark or Street Information in Route Directions: What Difference Does It Make? In Kuhn et al. (2003), pp. 362–374.

Woodruff, A. G. e Plaunt, C. (1994). GIPSY: Georeferenced Information Processing SYstem. *Journal of the American Society for Information Science*, 45(9):645–655.

Worboys, M. (2001). Nearness Relations in Environmental Space. *International Journal of Geographical Information Science*, 15(7):633–651.

Worboys, M.; Duckham, M. e Kulik, L. (2004). Commonsense Notions of Proximity and Direction in Environmental Space. *Spatial Cognition and Computation*, 4(4):285–312.

Zong, W.; Wu, D.; Sun, A.; Lim, E.-P. e Goh, D. H.-L. (2005). On Assigning Place Names to Geography Related Web Pages. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 354–362, Denver, CO, USA.