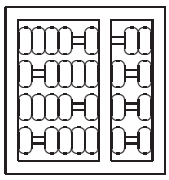

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Revisitando os desafios da recuperação de informação geográfica na Web

Lin Tzy Li *Ricardo da Silva Torres*

Technical Report - IC-09-18 - Relatório Técnico

May - 2009 - Maio

Revisitando os desafios da recuperação de informação geográfica na Web

Lin Tzy Li* Ricardo da Silva Torres†

Abstract

The geographic information ispart of people's daily life. There is a huge amount of information on the Web about or related to geographic entities and people are interested in localizing them on maps. Nevertheless, the conventional Web search engines, which are keywords-driven mechanisms, do not support queries involving spatial relationships between geographic entities. This paper revises the Geographic Information Retrieval (GIR) area and restates its research challenges and opportunities, based on a proposed architecture for executing Web queries involving spatial relationships and an initial implementation of that.

Resumo

Há uma grande quantidade de informação na Web sobre entidades geográficas e grande interesse em localizá-la em mapas. Entretanto, os mecanismos de busca na Web ainda não suportam em uma única ferramenta buscas que envolvam relações espaciais, pois em geral a consulta é processada levando-se em conta apenas as palavras-chaves usadas na consulta. Este artigo faz uma breve revisão da área de Recuperação de Informação Geográfica (GIR) e uma releitura de desafios e oportunidades de pesquisa da área a partir da proposta de uma arquitetura para buscas Web envolvendo relacionamento espacial entre entidades geográficas e uma implementação inicial dela.

1 Introdução

A informação geográfica pressupõe a existência de atributo relacionado à localização no espaço, como por exemplo uma coordenada geográfica ou relação direta ou indireta a algum objeto que possa ser localizado geograficamente. Isso pode ser desde um endereço completo, até referências como aeroporto de Cumbica, o que remete ao município de Guarulhos, próximo à cidade de São Paulo.

O que se procura na área de *Geographic Information Retrieval* (GIR) – Recuperação de Informação Geográfica – é tratar os novos desafios derivados da adição da variável geográfica na tradicional área de recuperação de informação. A área de GIR pode ser entendida

*Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP & Fundação CPqD, Campinas, SP.

†Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.

como uma extensão da área de *Information Retrieval* (IR) – Recuperação de Informação – incrementada com associações e dados sobre objetos geográficos.

A informação geográfica está presente direta ou indiretamente no dia-a-dia das pessoas e, desta forma, não é de se admirar que haja uma grande quantidade de informação na Web sobre entidades geográficas e grande interesse em localizá-la em mapas. Ferramentas como Google Maps e Google Earth vêm popularizando e atendendo, em parte, necessidades dos usuários Web por informação geoespacial.

Os serviços de buscas convencionais são baseados em casamento de palavras-chaves e em geral não levam em conta que estas palavras podem representar entidades geográficas que se relacionam espacialmente com outras entidades geográficas. Mesmo que não tenham sido citadas explicitamente na consulta [25], estes relacionamentos representam potencialmente uma informação relevante para o usuário.

Um exemplo de consulta que a maioria dos sistemas GIR existentes não suporta seria: "Quais são as páginas das prefeituras das cidades vizinhas a Campinas?". A dificuldade em se processar este tipo de consulta reside em combinar consultas tradicionais feitas em mecanismos de busca na Web com operadores espaciais, usualmente implementados em bancos de dados espaciais.

Este artigo apresenta desafios e oportunidades de pesquisa relacionados ao processamento de buscas Web envolvendo relacionamento espacial entre entidades geográficas. Primeiramente é dada uma visão geral sobre os conceitos da área de recuperação de informação geográfica, seguida da caracterização de desafios na área e de uma proposta de arquitetura para GIR. Por fim, é apresentado um mapeamento de novas oportunidades de pesquisa na área.

2 Visão Geral da área de Recuperação de Informação Geográfica (GIR)

A área de recuperação de informação geográfica tem foco na recuperação e indexação geoespacial da informação. Ela é uma área aplicada de pesquisa que combina pesquisa em sistema gerenciador de banco de dados (SGBD), interface humano-computador (IHC), sistema de informação geográfica (SIG/GIS), indexação, recuperação da informação (IR) e navegação (browsing) pela informação geo-referenciada [26], além da sua visualização espacial em um mapa.

2.1 Recuperação de Informação (IR)

Recuperação de informação lida com o desafio de se buscar informação sobre determinado assunto em que um usuário tem interesse. No entanto, o ser humano expressa suas necessidades em linguagem natural e está longe de ser preciso na formulação de suas consultas. Desta forma, uma das preocupações da área de IR consiste em interpretar a consulta formulada pelo usuário, buscar a informação armazenada em repositórios, selecioná-la conforme a sua relevância para o assunto de interesse do usuário, classificá-la (*rank*) e mostrar o conjunto resultado de forma adequada. Como a própria consulta envolve um grau de im-

precisão, o resultado retornado também contém uma margem de itens não relevantes. O objetivo principal para IR é maximizar os resultados relevantes e minimizar os irrelevantes.

Em linha geral, a efetividade do Recuperação de Informação é diretamente influenciada pela atividade do usuário (*user task*) e representação lógica (*logic view*) do documento adotada pelo sistema [2].

O processo de recuperação de informação que envolve o usuário (*user task*) se subdivide em duas partes [2]: o processo de formulação da consulta e o processo de refinamento (*browsing*) que o usuário efetua sobre os resultados retornados pelo sistema a fim de obter efetivamente os documentos desejados. É justamente nestes pontos que as pesquisas em recuperação de informação centradas no usuário focam, estudando o comportamento humano e as suas necessidades visando melhorar a modelagem, a organização e a execução de consultas no sistema. Já as pesquisas centradas no computador se preocupam principalmente em construir índices, processar as consultas dos usuários eficientemente e desenvolver algoritmos de classificação/ordenação (ranking) que melhorem a qualidade da resposta do sistema.

2.2 Arquitetura de um sistema de Recuperação de Informação (IR)

A Figura 1 ilustra uma arquitetura típica de um sistema de recuperação de informação.

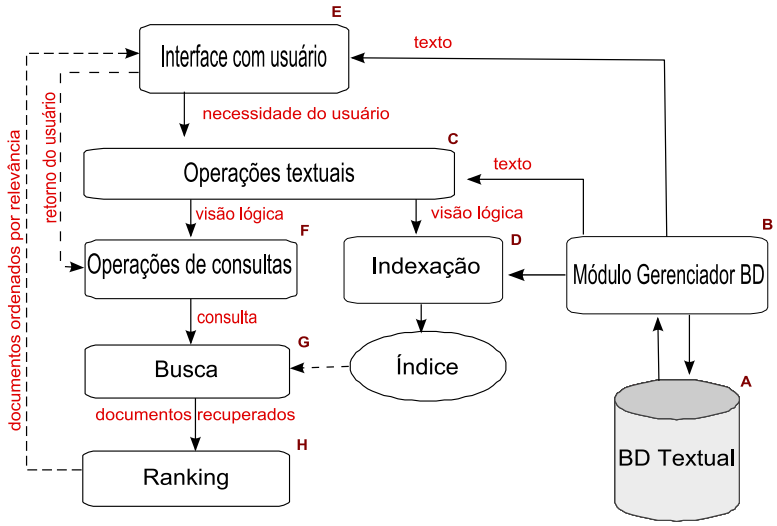

Figura 1: Processo de recuperação de informação (IR) [2].

Os principais módulos representados pela Figura 1 são:

Interface com usuário (E): Este módulo é responsável pelo recebimento das consultas formuladas pelo usuário e pela visualização dos resultados retornados pelo sistema de IR.

Operações textuais (C): Este módulo lida com formatos disponíveis para representar a informação e as propriedades do documento. Ele também é responsável pelo seu

pré-processamento – análise léxica, eliminação de termos irrelevantes (*stopwords*), identificação da raiz da palavra e sinônimos –, compressão de texto e agrupamento (clustering) de documentos.

Operações de consulta (F): Este módulo é encarregado de interpretar a consulta recebida. Além disso, cuida das interações subseqüentes, visando ao refinamento dos resultados.

Busca e Indexação (G e D): Este módulo se preocupa em recuperar a informação de forma mais eficiente, usando métodos de indexação, ténicas de casamento de padrão, consultas estruturadas e/ou consultas sobre índices comprimidos.

***Ranking* (H):** Este módulo ordena os documentos de acordo com a sua relevância para a necessidade do usuário.

Módulo gerenciador de BD (B): Este módulo define os documentos usados, o seu modelo de dados e as operações válidas. O módulo gerenciador de banco de dados constrói os índices dos textos para melhorar o desempenho da recuperação de informação. O espaço gasto com índices e o tempo necessário para sua criação são compensados pela diminuição do tempo de espera de resposta de um sistema de recuperação de informação.

O início do processamento de uma consulta na sistema de IR acima é disparado pelo o usuário a partir da especificação de uma consulta. Em seguida, o módulo de operações textuais do sistema processa e transforma a expressão da necessidade do usuário em algo cuja visão lógica seja da mesma natureza dos textos armazenados no sistema. A visão lógica da necessidade do usuário é submetida às operações de consulta (F) para que transforme a necessidade do usuário em uma consulta apropriada computacionalmente. O processamento da consulta utiliza os índices construídos e armazenados no sistema previamente. Os documentos recuperados são, então, classificados de acordo com a probabilidade de relevância deles conforme a necessidade expressa pelo usuário. Em seguida, são apresentados para avaliação e determinação do subconjunto de documentos que realmente seja de interesse. Neste ponto, a indicação do subconjunto de interesse pelo usuário é uma retro-alimentação para o sistema refinar os resultados a serem apresentados em uma próxima iteração.

2.3 Propriedades da informação geo-referenciada

Uma informação pode ser considerada geo-referenciada quando ela tem uma coordenada associada, ou quando a informação faz referência a alguma "entidade geográfica", por exemplo, nomes de lugares ou frases que remetem a lugares [37]. A associação de determinado item de uma coleção a uma ou mais regiões na superfície terrestre é denominada por [14] como *footprint*. Jones [24] define a ação de associar um *footprint* a uma referência geográfica de **geo-codificação**. Já a ação de reconhecer uma referência geográfica é denominada como **geo-parsing** (análise sintática).

Em GIR, é necessário que a coleção de dados que referencia direta ou indiretamente lugares seja traduzida em seu *footprint* para que possa ser indexado espacialmente, ou

seja, passar por geo-parsing e então por geo-codificação. No entanto, alguns desafios são observados [37] neste processo:

- referências a lugares homônimos. Por exemplo, Nova York designa uma cidade no Maranhão ou um estado nos EUA, assim como Luis Eduardo Magalhães pode ser nome de um aeroporto, praça ou cidade na Bahia;
- lugares citados em textos mudam conforme contexto histórico, cultural e costumes populares em que estes texto são produzidos. Por exemplo, "200 km do norte da capital da Rússia" tem o problema da Rússia ter tido outras capitais ao longo da história;
- nomes de lugares mudam ao decorrer do tempo;
- extensão geográfica de um local muda com o tempo;
- fronteiras podem não ser claras;
- mesmo lugar pode aparecer escrito de diversas formas em diferentes textos, seja por erro, língua ou existência legal de mais de uma forma válida de escrita;
- ambigüidades por causa de referências feitas relativamente a um lugar, com pseudônimos, ou dentro de contextos específicos;
- referências indiretas. Por exemplo, Rodovia Fernão Dias remete aos estados de São Paulo e Minas Gerais, assim como falar de Cristo Redentor remete à cidade do Rio de Janeiro.

2.4 Ferramentas de geo-referenciamento

Gazetteers, Thesauri e Ontologias [5, 4] constituem técnicas comumente utilizadas para contornar as dificuldades enumeradas na seção anterior ao se fazer o geo-parsing e a geo-codificação.

2.4.1 Ontologias

Ontologia é um modelo de objetos, taxonomias e esquemas [5] e provê um conjunto de conceitos e termos para descrever um domínio e, portanto, uma estrutura sobre o qual uma base de conhecimento pode ser construída.

Ontologias são usadas como solução em vários domínios para representar o conhecimento, pela possibilidade de explicitar e especificar as semânticas e as relações do domínio. Uma das qualidades de ontologias é a flexibilidade de seu reuso e compartilhamento, além da possibilidade de acomodar uma variedade de termos descritivos [31]. Elas podem ser usadas para reconhecimento e extração de evidências geo-espaciais e precisamente neste contexto são denominadas de *ontologias geográficas* [5] ou *geo-ontologias* [35].

2.4.2 Gazetteers e Thesauri

Gazetteer é um dicionário de nomes geográficos cujos componentes principais são: nome e suas variantes, a localização e categoria do lugar, ajudando a responder questões do tipo "onde fica esse lugar?" e "o que há nesse lugar?" [22, 5]. Ele possui também informações descritivas dos lugares, podendo ser usado para associar coordenadas geográficas ao nome de um lugar [5].

Embora os Gazetteers contenham mais informação sobre determinado local geográfico identificado por um texto, eles não representam nenhuma relação semântica (por exemplo, sinônimo e 'hyponymy') ou espacial (por exemplo, vizinhança) entre lugares listados. Ao contrário do thesaurus, que enfatiza a relação espacial entre os lugares em detrimento à localização exata em termos de coordenadas.

Um thesaurus é uma lista de termos estruturada e definida que padroniza as palavras usadas com índices, ou seja, é um vocabulário formalmente organizado de tal forma que as relações entre conceitos são explicitadas [6]. Por exemplo, o *Getty Thesaurus of Geographic Names* [36] organiza o lugar por sua relação espacial e por áreas administrativas; informa as várias versões de nome que um lugar pode ter; informa suas coordenadas geográficas; e dá suporte a lugares com nomes similares usando ontologias [37].

2.5 Relacionamentos espaciais

Borges [5] agrupa os relacionamentos espaciais, as posições relativas entre objetos, em três categorias:

topológicos: estes relacionamentos indicam as relações de conectividade e não incluem medidas e direção, mas propriedades como adjacência, conectividade e relação de *contém* e *estar contido*. Egenhofer [12] classifica os relacionamentos topológicos entre dois objetos bidimensionais em: *disjunto*, *encontra*, *sobrepõe*, *contém*, *cobre*, *dentro*, *coberto por* e *igual*. Em contrapartida, Clementini [10] resume-os em: *disjunto*, *dentro*, *toca*, *cruza* e *se sobrepõe* (Figura 2).

métricos: os relacionamentos métricos expressam propriedades espaciais mensuráveis de forma quantitativa, como área, distância, comprimento e perímetro.

direcionais: estes relacionamentos expressam orientação – por exemplo os pontos cardinais norte, sul, leste e oeste – e ordem – por exemplo acima, abaixo, em frente.

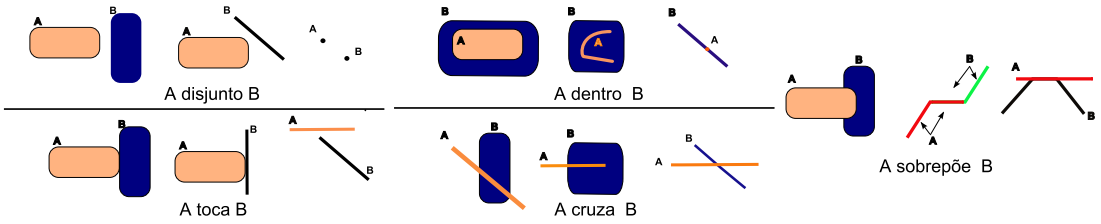

Figura 2: Exemplos de relacionamento topológicos [7].

2.6 Consultas espaciais

Consultas espaciais são consultas sobre as relações espaciais de dois objetos localizados em um espaço bem definido com coordenada geográfica ou não. Segundo Larson [26], as relações espaciais podem ser geométricas ou topológicas, sendo que a primeira inclui relação de distância e direção. Por exemplo, as coordenadas em latitude e longitude de Nova York (4040'N 7358'W) e de Chicago (4152'N 8737'W) nos indicam direção e a distância entre as cidades que podem ser calculadas a partir dessas coordenadas.

Segundo Larson [26] as consultas espaciais podem ser classificadas em:

- **por ponto em um polígono**: tenta responder consultas do tipo "O que há no ponto (x, y) do sistema de coordenada corrente?";
- **por regiões**: quando dada uma região delimitada por um polígono ou linha, tenta-se encontrar algo que esteja contido nele, adjacente a ele, ou que se sobreponha à sua área. Por exemplo, "Quais áreas têm intersecção com uma dada área escolhida?";
- **por distância e zona de buffer**: consiste em encontrar algo que está dentro de uma distância fixa de um objeto, seja uma linha, um ponto ou um polígono. Um exemplo de consulta deste tipo seria: "Quais são as cidades que estão a 50 km dos limites das cidade de Campinas?";
- **por caminhos**: é uma consulta que envolve uma estrutura de rede formada por segmentos de linha conectados, como é o caso de rede elétrica, canos de água ou gás, vias de transportes, etc. Consultas tradicionais são as de caminho mais curto entre dois pontos da rede. No entanto, consultas que envolvam variáveis diferentes de distância e direção podem ser mais complicadas (por exemplo, "Qual o caminho mais rápido de Campinas a Santo André?");
- **multimídia**: são as consultas que congregam informação de vários tipos de dados (textual, imagem, geográfico). Por exemplo, consulta do tipo "Quais rios que atravessam estados que possuem cidades cujos nomes contenha *Paulo* e possuem peixes similares àquele encontrado em uma dada imagem de entrada?".

2.7 Arquitetura de um sistema de GIR

A Figura 3 ilustra a arquitetura típica de um sistema de recuperação geográfica. Como pode ser observado, alguns módulos foram adicionados (área delimitada com pontilhado) e outros alterados (destacado em laranja) em relação à Figura 1.

Geo-coding (K): o geo-codificador de documento extrai a referência geográfica (footprint) de determinado documento com base em seu conteúdo.

Geo-parsing (J): módulo desambiguador que, a partir de ontologias e dados semânticos, uniformiza os termos geográficos ambíguos e similares semanticamente.

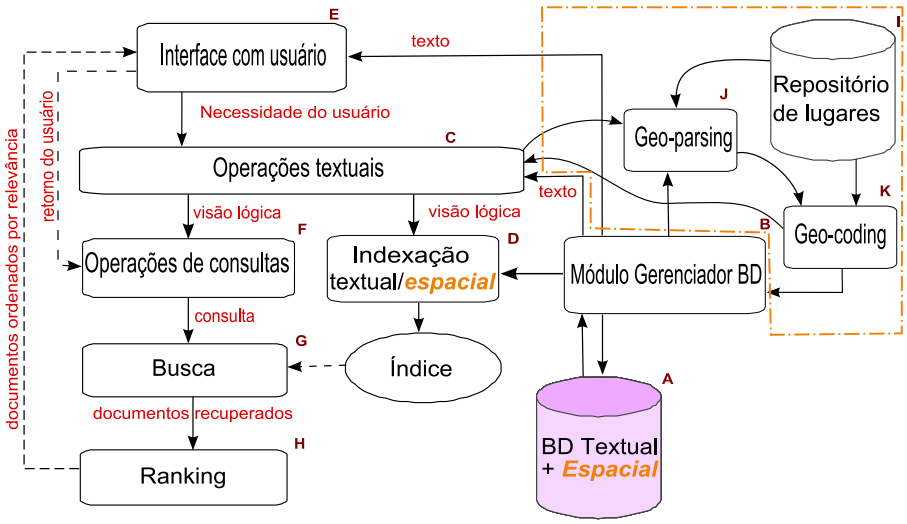

Figura 3: Arquitetura de um sistema de recuperação de informação geográfica.

Banco de dados espacial: base de dados de lugares geo-referenciado que é usada para ajudar a atribuir coordenadas geográficas a um conteúdo com base na sua referência geográfica. Exemplos de bases usadas são os gazetteers e, atualmente, até mesmo referências encontradas em páginas Web na internet [4] e outros documentos relacionados que tenham sido previamente geocodificados.

Busca e Indexação espaciais (G e D): se encarrega também em prover meios para as consultas espaciais serem mais eficientes, ou seja, prover métodos de acessos (SAM) eficientes usando as coordenadas geográficas associadas aos documentos como índices. O processamento de consultas espaciais usa técnicas de geometria computacional para descobrir as relações entre os objetos espaciais, representados por geometrias como ponto, linha ou polígono. Entre os esquemas usados para os índices espaciais citam-se: Linear Quadtree, Space-filling curves, árvore Z-Ordering, árvore R, R* e R+ [33]. O módulo de busca trata também da ordem em que uma busca deve ser feita quando as consultas são simultaneamente alfanuméricas e espaciais. Uma das preocupações, por exemplo, é como a ordem de execução dos tipos de consultas pode afetar o desempenho do sistema de busca [9].

Navegação espacial (spatial browsing): é o módulo de interface com usuário (E). Pode prover a visualização, em mapa, da localização do documento ou do local sobre o qual o documento versa, bem como pode oferecer alguma forma do usuário refinar sua consulta via mapa.

Operações de consulta e Ranking (F e H): estes módulos se preocupam em como tratar a introdução da variável espacial nas consultas do usuário, como por exemplo traduzir as palavras com significado geográfico (objetos ou operadores) em uma liguagem de sistema, e como alterar o algoritmos de *ranking* dos resultados de forma a retornar

somente aqueles que são relevantes ao usuário.

3 Caracterização de novos desafios na área de Recuperação de Informação Geográfica

Nesta seção será usado um estudo de caso para caracterizar problemas relacionados ao processamento de consultas Web que consideram relacionamento espacial entre objetos geográficos. Primeiramente será dado uma visão geral das possíveis aplicações que se beneficiariam deste tipo de consulta.

3.1 Aplicações

Exemplos de possíveis aplicações são aquelas relacionadas à busca de documentos sobre lugares de interesse (aplicações, por exemplo, na área de turismo e projetos de engenharia). Considere os cenários abaixo, sendo que as palavras em negrito estão associadas a relacionamentos espaciais entre entidades geográficas:

- Você mora em Curitiba e gostaria de prestar concursos públicos para trabalhar na prefeitura de cidades **vizinhas**. Assim gostaria de acessar as páginas das prefeituras e procurar por editais em aberto.

- Você está indo para uma visita turística à Curitiba e região. Você tem restrição financeira e sabe que hotéis da capital são mais caros, por isso gostaria de procurar hotéis das **redondezas** para ficar, mas ainda viabilizando o seu roteiro de visita a Curitiba.

- Você tem interesse em viajar pelo estado São Paulo e quer aproveitar a viagem para fazer uma pesquisa sobre vilas e cidades que ficam **perto** do Rio Tietê. Assim, seria interessante encontrar as páginas das concessionárias das estradas que **cruzam** o rio para contatá-las, a fim de propor patrocínio ou algum trabalho conjunto de interesse.

- Você está indo para uma conferência em Barcelona que será realizada em um local próximo a uma estação de metrô e quer aproveitar para conhecer a cidade. Neste caso, seria interessante se hospedar em hotéis **próximos** a qualquer estação de metrô a fim de facilitar sua locomoção pela cidade.

- Você participa de um projeto de inclusão digital que experimentalmente ligará com um cabo de comunicação de última geração as cidades de Campinas e Peruíbe, que fica no litoral, ao **Sul** de Campinas. Você imagina que mais cidades poderiam se beneficiar com esta mesma ligação. Por isso, tem a idéia de entrar em contato com as prefeituras de outras cidades ao **Sul** de Campinas, cujas regiões potencialmente **serão cruzadas** pelo cabo de transmissão.

3.2 Estudo de caso

Em uma enquete informal, envolvendo 15 pessoas com diversos níveis de conhecimento de uso do computador, perguntou-se como fariam para encontrar páginas Web, considerando o seguinte cenário: "Quais são as páginas das prefeituras das cidades vizinhas (até 50km) da cidade X?". Várias soluções foram apresentadas:

Solução 1: Submeteria ao *Google Search* o nome da (micro) região da cidade, por exemplo, "triângulo mineiro", "circuito das águas" mais o objetos de interesse: *prefeitura "circuito das águas"* .

Solução 2: Submeteria ao Google Search as palavras-chaves *prefeitura região X*.

Solução 3: Submeteria ao Google Search as palavras-chaves *cidades vizinhas X* e depois com a lista da cidades em mãos, procuraria pelas páginas da prefeitura de cada cidade.

Solução 4: Usaria o Google Maps buscando pela cidade X, inspecionaria visualmente o mapa para listar as cidades vizinhas ou as próximas e faria busca no Google pelos nomes da cidades com chaves *<nome da cidade> prefeitura*.

Solução 5: Submeteria ao google search as palavras-chaves *prefeitura próxima <cidade X> 50 km*.

Solução 6: Procuraria uma lista de cidades do estado via busca Web ou iria no site do governo do estado para obtê-la. Com sorte esta página já poderia conter os links para as páginas de prefeituras. Com a lista em mãos, procuraria por uma tabela de distâncias na Web, para finalmente fazer a consulta na google search com *<nome da cidade> prefeitura*.

Solução 7: Primeiro submeteria ao Google Search *cidades distâncias X* para recuperar as cidades de interesse. Em seguida, para cada cidade da lista, buscaria por *prefeitura <cidade>*.

Solução 8: Considerando que sejam cidades de São Paulo, pegaria a lista das cidades da região em que X se insere, depois iria ao endereço www.<cidade>.sp.gov.br, substituindo <cidade> pelo nome da cidade, pois este é o padrão que o endereço das páginas da prefeitura costumam seguir.

Solução 9: Submeteria ao Google Search as palavras-chaves *Sites prefeituras Campinas SP região*.

Solução 10: Visitaria a página da cidade no Wikipedia que costuma ter a informação de cidades vizinhas.

Muitas destas soluções apresentam mais de um passo para se responder à consulta desejada. Em geral porque estes usuários (que podem ser considerados mais experientes em buscas geográficas) já sabiam que as ferramentas de busca Web atuais não respondem tão bem as consultas deste tipo.

O uso de apenas o nome da cidade e a relação espacial desejada não é suficiente para que a informação relevante seja recuperada, pois a máquina de busca apenas tentará casar as palavras-chaves usadas. Para usuários que estão costumados a fazer este tipo de consulta, é comum tentar reescrevê-la de modo que a ferramenta de busca Web retorne resultados relevantes.

Tomando como exemplo a consulta feita, pode-se dizer que, de um modo geral, ela foi fatorada da seguinte forma. Primeiramente é usada alguma informação que ajude a transformar a consulta geográfica em uma consulta por palavras-chaves: (a) se valendo de conhecimentos prévios, associando a cidade a uma região que a englobe (estado, região) que já é de seu conhecimento (Solução 1), ou indo direto à página da prefeitura, pois se sabe a estrutura URL das páginas da prefeitura (Solução 8); (b) ou visitando páginas previamente conhecidas que poderiam possuir a lista das cidades próximas ou a tabela de distância entre cidades (Solução 10); (c) ou submetendo outras palavras ao serviço de busca para retornar a lista das cidades vizinhas ou próximas (Soluções 3, 6 e 7); (d) ou usando o serviço de localização de mapas da cidade-referência da consulta e visualmente discernir e manualmente listar as cidades que possui a relação geográfica desejada (Solução 4).

No passo seguinte, o usuário monta a consulta ou as consultas com as palavras-chaves que terão mais chance de retornar resultados relevantes, levando-se em conta a lista de cidades alvos que foram definidas pela relação espacial da consulta inicial.

Com este cenário em mente, propõe-se uma arquitetura para enriquecer a busca Web tradicional adicionando consultas geográficas com o auxílio de banco de dados espacial. A proposta é que o usuário expresse diretamente sua consulta geográfica e o sistema expanda esta consulta, envie-a a máquinas de buscas existentes, combine os resultados e retorne ao usuário os resultados ordenados por sua relevância. Em seguida, alguns desafios e oportunidades de pesquisa relacionados à implementação dessa arquitetura são vislumbrados.

3.3 Arquitetura proposta

A arquitetura proposta neste trabalho é um modelo de três camadas, conforme ilustrado na Figura 4. Na camada de visualização tem-se a interface humano-computador para definição da consulta pelo usuário, o retorno dos resultados e o refinamento da consulta. Prevê-se a possibilidade de usar APIs externas para ajudar na exibição de informação extraída da Web, por exemplo o Google Maps API [19], que são providas externamente ao sistema para ajudar o desenvolvedor adicionar, em suas páginas, funcionalidades providas por outros sites.

Na camada de processamento da entrada, encontra-se o módulo responsável pelo *geo-parsing* de termos usados na consulta, o geocodificador da consulta, o módulo de expansão de consulta, o gerenciador de máquinas de buscas, o refinador (feedback) de consultas e o módulo de ranking por relevância. A máquina de busca pode repassar a busca para várias outras existentes na Web de forma que o resultado do sistema será a combinação dos resultados retornados pelas diversas máquinas de buscas.

Por fim, a camada de dados é composta por repositórios locais e os distribuídos pela Web. Estes repositórios consistem de dados, ontologias e thesauri para desambiguar termos ou expandir a consulta do usuário. Os repositórios remotos podem conter também outras

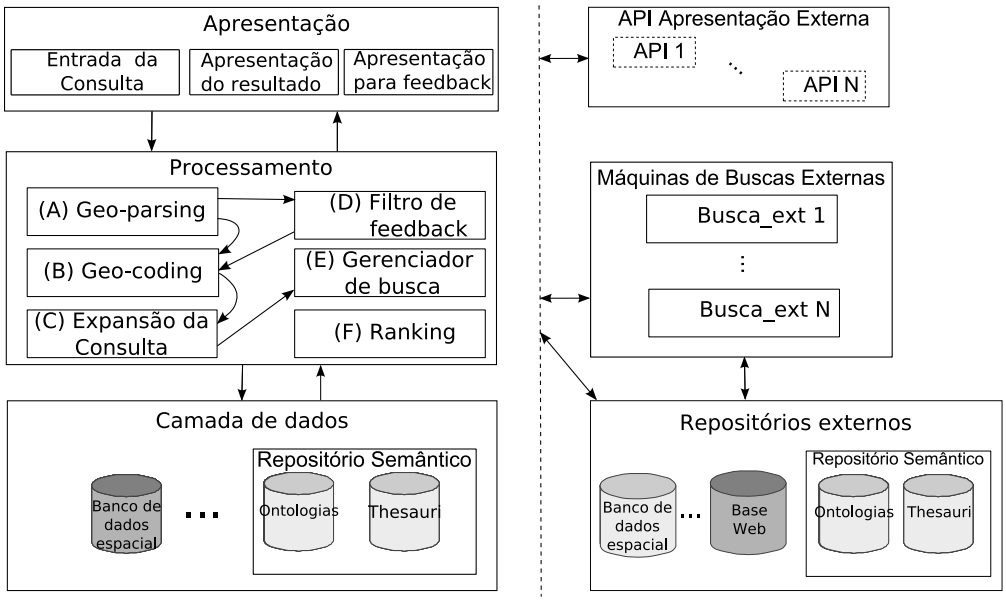

Figura 4: Arquitetura para recuperação informação geográfica na Web.

ontologias e thesauri e incluem, ainda, os documentos disponíveis na Web.

Um típico cenário de uso consiste nas seguintes etapas: o usuário especifica sua consulta; o sistema reconhece e desambigua os termos que se referem a objetos geográficos da consulta, por exemplo, os nomes de lugares homônimos ou que se referem a mais de um objeto; o sistema pode pedir para o usuário filtrar e indicar o sentido ou contexto correto dos termos a serem usados na consulta, passando o controle para a interface. O usuário indica na interface o sentido e o contexto, então o sistema geocodifica os elementos de referência da consulta geográfica e prepara a consulta para ser enviada para o gerenciador de máquinas de busca. O resultado da busca passa por um ranking de relevância antes de ser apresentado para o usuário. Com a visualização do resultado, o usuário pode querer filtrar mais ainda o resultado, realimentando o sistema com novos critérios para uma nova busca.

4 Protótipo

Parte da arquitetura proposta na Figura 4 foi implementada em um protótipo. Os módulos implementados foram: Entrada da Consulta, Apresentação do resultado, Geocodificação do objeto de referência da consulta (B), Expansão da Consulta (C), Busca (E), Banco de dados espacial e uso de API de apresentação.

Buscas envolvendo relacionamentos espaciais foram implementadas por meio de consultas enviadas a um banco de dados espacial. O banco de dados espacial foi carregado com dados vetoriais obtidos do site do IBGE [23] como cidades, estados, rios, rodovia federal e ferrovia do Brasil.

A consulta é estruturada em uma interface Web (Figura 5) com campos fixos. Na pri-

meira caixa de seleção, o usuário indica o tipo de informação (por exemplo, páginas de prefeitura) de interesse e o tipo de objeto geográfico ao qual esta informação se relaciona (por exemplo, cidade), que será chamado de objeto-alvo. Em seguida, escolhe-se a relação espacial (por exemplo, vizinho) que estes objetos-alvos devem ter com um objeto de referência (objeto-referência), que o usuário especifica também o tipo do objeto espacial e o caracteriza (por exemplo, cidade X).

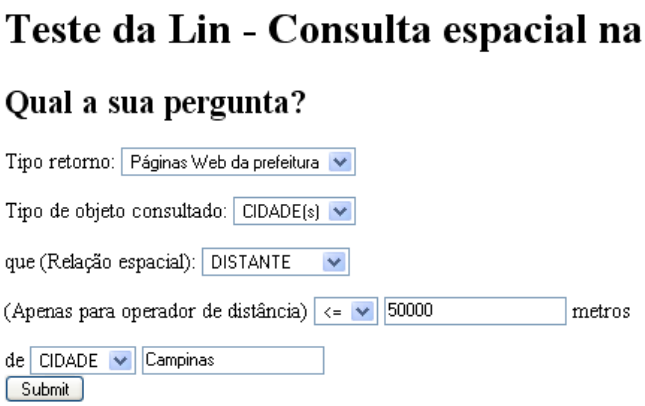

Figura 5: Interface para especificação de consultas envolvendo relacionamento espacial entre objetos geográficos.

No processamento da consulta, tendo o objeto-referência bem caracterizado ele pode ser usado na consulta geográfica equivalente fornecida pelo BDE para busca da lista de objetos-alvos. Com a lista de objetos em mãos, expande-se a consulta espacial de entrada e envia-se a nova consulta para uma máquina de busca Web (no caso, Google). O resultado da busca é mostrado numa página Web em que se agrega os resultados retornados na busca e a localização espacial dos objetos-alvos no mapa (Figura 6). Desta forma, o usuário consegue recuperar a informação de interesse em apenas um passo.

Este protótipo foi implementado usando a linguagem de programação Javascript e Python [30] sob o framework para aplicações Web Django [11]. A máquina de busca Web foi provida pelo Google AJAX Search API [16] e a exibição da localização no mapa dos objetos-alvos foi fornecida pelo Google Maps API [17]. Como repositório de dados espacial, foi adotado o PostgreSQL [29] com extensão espacial PostGIS [32] e nele foram carregados vários dados vetoriais obtidos do site do IBGE [23] como cidades, estados, rios, rodovia federal e ferrovia do Brasil.

5 Novos desafios e oportunidades de pesquisa em GIR

A especificação e a implementação da arquitetura proposta na seção anterior requer lidar com vários desafios de pesquisa. Nesta seção, alguns desses desafios são discutidos, levando-se em consideração as três principais camadas da arquitetura: apresentação, processamento e de dados.

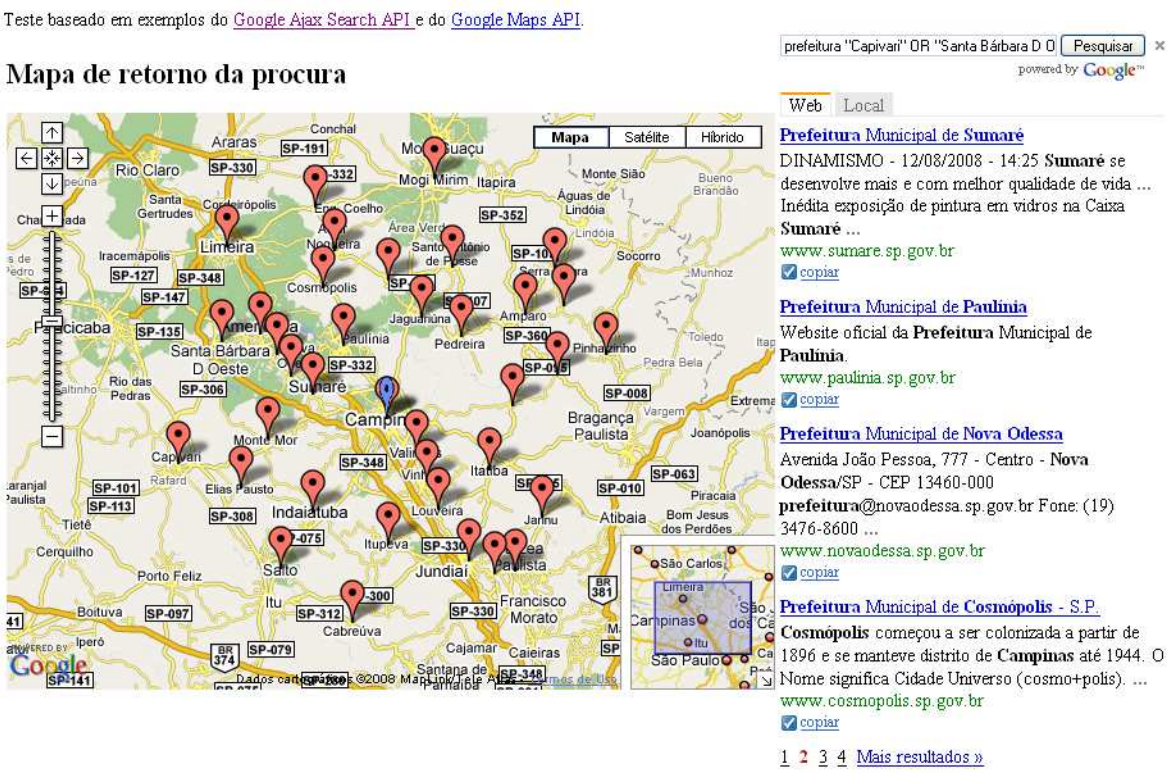

Figura 6: Resultado da consulta "Quais são as páginas das prefeituras das cidades próximas (até 50 km) da cidade de Campinas?".

5.1 Camada de apresentação

A interação humano-computador mais primitiva ainda exige que o usuário formule a sua consulta de forma estruturada próximas de linguagem de consulta de banco de dados (por exemplo SQL). Como a maioria dos usuários não conhece suficientemente essa linguagem estruturada, eles não conseguem expressar completamente suas necessidades e, conseqüentemente, a informação recuperada não satisfaz às suas expectativas de fato. Em se tratando do usuário precisar traduzir sua noção espacial em palavras na consulta, introduz-se mais uma complexidade e indireção ao problema. Por outro lado, a consulta precisa mesmo ser expressa apenas por meio de palavras?

A dificuldade em oferecer uma interface em que o usuário possa expressar sua necessidade em uma linguagem natural, por exemplo, está em problemas que os pesquisadores da área de processamento de linguagem natural vem tentando solucionar há décadas: ambigüidades, imprecisão e dependência de contexto na linguagem humana. Este desafio se torna maior ainda ao se adicionar variáveis espaciais, pois o ser humano refere-se a lugares de forma imprecisa, sem mencionar a relação de temporalidade dos lugares, conforme discutido na seção 2.3.

Nas buscas locais do Google (Google Local Search) [18], um conjunto de páginas geo-

codificas são recuperadas em consultas no Google Maps e, portanto, passíveis de serem localizadas em mapa. No entanto, ainda há poucas páginas geocodificadas, tanto que com as mesmas palavras-chave usada na busca local, se usada na busca Web usualmente retorna mais resultados. Além disso, o próprio Google Maps oferece alguma consulta do tipo "perto de" um ponto selecionado no mapa usando o Local Search. Neste caso, uma possível estratégia seria usar uma interface desse tipo, agregando consultas que envolvam outras outras relações espaciais.

Há ainda vários desafios a serem tratados na camada de apresentação com relação à forma pela qual o usuário poderia expressar a sua consulta, o resultado poderia ser apresentado, ou o usuário poderia interagir com o sistema indicando quais resultados são realmente relevantes de modo que o sistema possa aprender a refinar os resultados que serão apresentados em uma próxima iteração.

5.2 Camada de processamento

Já na camada de processamento, há o desafio de desambiguar nomes de lugares, caso tenha sido usado um nome que é comum a vários lugares e objetos, ou que possui um nome alternativo, por exemplo. Neste caso, nomes similares podem ser apresentados para filtragem do usuário; uma nova consulta, considerando o feedback do usuário, é enviada ao sistema de busca.

Por outro lado, como foi proposto submeter a consulta a várias máquinas de buscas existentes, entre os desafios estão os de combinar os resultados de várias fontes, fazer um ranking de relevância deles e tratar o feedback [39, 20] do usuário com relação à relevância dos resultados apresentados e interagir com os diversas máquinas de busca.

Supondo que se ofereça uma interface para o usuário expressar sua necessidade através da linguagem semi-estruturada ou natural, o desafio será como identificar e manipular referências a lugares nas consultas Web [8, 34] e lidar com imprecisões dessas referências [27, 15]. Como forma de tentar considerar estas questões, há trabalhos para caracterizar as necessidades do usuário quanto à informação geográfica [21].

Se a base de conhecimento geográfica estiver devidamente montada e geocodificada, há ainda o desafio de se processar eficientemente a consulta em máquinas de buscas geográficas na Web [9], considerando a quantidade de dados que a Web representa.

Outro desafio é produzir algoritmos eficazes para determinar a relevância do documento ou objeto frente às necessidades expressas pelo usuário, por exemplo utilizando técnicas de aprendizado [38, 13].

5.3 Camada de dados

Considerando que a própria Web pode ser vista como um grande repositório de dados, então a criação de uma base de conhecimento geográfico de forma automática com base em informação disponível na Web já constitui um desafio e tanto. Neste caso, lida-se com informação inconsistente [28] e com o desafio de identificar e de geocodificar dados textuais não-estruturados encontrados nas páginas Web [4, 1, 3].

6 Conclusão

Este artigo apresentou uma breve revisão da área de recuperação de informação geográfica, buscando caracterizar alguns dos principais desafios na área. A percepção é que os mecanismos de busca na Web ainda não suportam, em uma única ferramenta, buscas que envolvam relacionamentos espaciais entre entidades geográficas, pois em geral a consulta é processada levando-se em conta apenas as palavras-chaves usadas na consulta.

Visando prover consultas geográfica na Web usando mecanismos de busca existentes e banco de dados espacial, foi proposta uma arquitetura. A partir da arquitetura proposta, foi feita uma releitura, uma implementação de um protótipo inicial e identificação de novo desafios e oportunidades de pesquisa na área de recuperação de informação geográfica.

7 Agradecimentos

Este trabalho contou com o apoio de FAPESP, CNPq, CAPES e CPqD.

Referências

[1] Mirna Adriani and Monica Lestari Paramita. Identifying location in indonesian documents for geographic information retrieval. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 19–24, Lisbon, Portugal, 2007.

[2] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., New York, NY, USA, 1999.

[3] Andre Blessing, Reinhard Kuntz, and Hinrich Sch utze. Towards a context model driven german geo-tagging system. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 25–30, Lisbon, Portugal, 2007.

[4] Karla A. V. Borges, Alberto H. F. Laender, Claudia B. Medeiros, and Jr Clodoveu A. Davis. Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 31–36, Lisbon, Portugal, 2007.

[5] Karla Albuquerque Vasconcelos Borges. *Uso de uma ontologia de lugar urbano para reconhecimento e extração de evidências geoespaciais na Web*. Doctoral thesis, UFMG - Universidade Federal de Minas Gerais, 2006.

[6] Daniela F. Brauner, Marco A. Casanova, and Ruy L. Milidiú. Towards gazetteer integration through an instance-based thesauri mapping approach. In *Brazilian Symposium on GeoInformatics*, Campos do Jordão, SP, Brazil, 2006. S6 - Distributed GIS / GIS and the Internet.

[7] Gilberto Câmara, Marco A. Casanova, Andréa S. Hemerly, Geovane C MAGALH AES, and Cláudia M. B. Medeiros. Anatomia de sistemas de informação geográfica. In

10a. Escola de Computação, page 197, Campinas, 1996. Instituto de Computação - UNICAMP.

[8] Nuno Cardoso and Mário J. Silva. Query expansion through geographical feature types. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 55–60, Lisbon, Portugal, 2007.

[9] Yen-Yu Chen, Torsten Suel, and Alexander Markowetz. Efficient query processing in geographic web search engines. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 277–288, Chicago, IL, USA, 2006.

[10] Eliseo Clementini, Paolino Di Felice, and Peter van Oosterom. A small set of formal topological relationships suitable for end-user interaction. In *Symposium on Advances in Spatial Databases*, pages 277–295. Springer-Verlag, 1993.

[11] Django Software Foundation. Django: The web framework for perfectionists with deadlines. http://www.djangoproject.com/.

[12] MAX J. EGENHOFER. Query processing in spatial-query-by-sketch. *Journal of Visual Languages & Computing*, 8:403–424, August 1997.

[13] Weiguo Fan, Praveen Pathak, and Linda Wallace. Nonlinear ranking function representations in genetic programming-based ranking discovery for personalized search. *Decision Support Systems*, 42(3):1338 – 1349, 2006.

[14] J. Frew, M. Freeston, N. Freitas, L. Hill, G. Janée, K. Lovette, R. Nideffer, T. Smith, and Q. Zheng. The alexandria digital library architecture. *International Journal on Digital Libraries*, 2:259–268, May 2000.

[15] Gaihua Fu, Christopher B. Jones, and Alia I. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, Lecture Notes in Computer Science, pages 1466–1482. Springer Berlin / Heidelberg, 2005.

[16] Google. Google ajax search api - google code. http://code.google.com/apis/ajaxsearch/.

[17] Google. Google maps api reference - google maps api - google code. http://code.google.com/apis/maps/documentation/reference.html.

[18] Google. Local search examples - google ajax search api - google code. http://code.google.com/apis/ajaxsearch/local.html.

[19] Google. Map basics - google maps api - google code. http://code.google.com/apis/maps/documentation/introduction.html.

[20] Daqing He. A study of self-organizing map in interactive relevance feedback. In *ITNG '06: Proceedings of the Third International Conference on Information Technology:*

New Generations, pages 394–401, Washington, DC, USA, 2006. IEEE Computer Society.

[21] Andreas Henrich and Volker Luedecke. Characteristics of geographic information needs. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 1–6, Lisbon, Portugal, 2007.

[22] Linda Hill. Core elements of digital gazetteers: Placenames, categories, and footprints. In *Research and Advanced Technology for Digital Libraries*, volume 1923/2000 of *Lecture Notes in Computer Science*, pages 280–290. Springer Berlin / Heidelberg, 2000.

[23] IBGE. Mapas interativos - ibge. http://www.ibge.gov.br/mapas/.

[24] Christopher Jones. Geographic information retrieval, November 2006. "See www.geospirit.org for information on SPIRIT project, the contributing partners, and downloads of articles and project deliverables.".

[25] Christopher B. Jones, Alia I. Abdelmoty, David Finch, Gaihua Fu, and Subodh Vaid. The spirit spatial search engine: Architecture, ontologies and spatial indexing. In *Geographic Information Science*, Lecture Notes in Computer Science, pages 125–139. Springer Berlin / Heidelberg, 2004.

[26] Ray R. Larson. Geographic information retrieval and spatial browsing. In *32nd Clinic on Library Applications of Data Processing*, pages 81–124, University of Illinois, Urbana-Champaign, April 1995. Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

[27] Robert C. Pasley, Paul D. Clough, and Mark Sanderson. Geo-tagging for imprecise regions of different sizes. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 77–82, Lisbon, Portugal, 2007.

[28] Adrian Popescu, Gregory Grefenstette, and Pierre Alain Mo ellic. Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 85–93, Pittsburgh PA, PA, USA, 2008.

[29] PostgreSQL. PostgreSQL: The world's most advanced open source database. http://www.postgresql.org/.

[30] Python Software Foundation. Python programming language – official website. http://www.python.org/.

[31] Jian Qin and Stephen Paling. Converting a controlled vocabulary into an ontology: the case of gem. *Information Research: an international electronic journal*, 6:94, 2001. published four times a year by Professor Tom Wilson of the Department of Information Studies, University of Sheffield.

[32] Refractions Research. PostGIS: Home. http://postgis.refractions.net/.

[33] Philippe Rigaux, Michel O. Scholl, Agnes Voisard, and Jim Gray. *Spatial Databases: With Application to GIS*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, California, CA (EUA), 2002.

[34] Mark Sanderson and Yu Han. Search words and geography. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 13–14, Lisbon, Portugal, 2007.

[35] Diana Santos and Marcirio Silveira Chaves. The place of place in geographical ir. In *3rd Workshop on Geographic Information Retrieval, SIGIR'06*, pages 5–8, Seattle, August 2006. Department of Geography, University of Zurich.

[36] J. Paul Getty Trust. Getty thesaurus of geographic names (research at the getty). http://www.getty.edu/.

[37] Øyvind Vestavik. Geographic information retrieval: An overview. In *Internal Doctoral Conference*, page 7, IDI, NTNU, Norway, 2003.

[38] Jun Xu, Tie-Yan Liu, Min Lu, Hang Li, and Wei-Ying Ma. Directly optimizing evaluation measures in learning to rank. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 107–114, New York, NY, USA, 2008. ACM.

[39] Zhao Xu, Xiaowei Xu, and Volker Tresp. A hybrid relevancefeedback approach to text retrieval. In *In Proceedings of the 25th European Conference on Information Retrieval Research, Lecture Notes in Computer Science*, pages 281–293. Springer-Verlag, 2003.