Geocodificação de endereços urbanos com indicação de qualidade

Douglas Martins[1], Clodoveu A. Davis Jr.[1], Frederico T. Fonseca[2]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Av. Presidente Antônio Carlos, 6627 – 31270-010 – Belo Horizonte – MG

[2]College of Information Sciences and Technology – The Pennsylvania State University
332 IST Building – 16802-6823 – University Park – PA – USA

[dougmf,clodoveu]@dcc.ufmg.br, ffonseca@ist.psu.edu

Abstract. *Urban addresses are one of the most important ways to express a geographic location in cities. Many conventional information systems have attributes for addresses in order to include an indirect reference to space. Obtaining coordinates from addresses is one of the most important geocoding methods. Such activity is hindered by frequent variations in the addresses, such as abbreviations and missing components. This paper presents a geocoding method for urban addresses, in which address fragments are recognized from the input and a reference geographic database is searched for matching addresses for the corresponding coordinates. Output includes a geographic certainty indicator, which informs the expected quality of the results. An experimental evaluation of the method is presented.*

Resumo. *Endereços urbanos são uma das principais formas de expressão da localização geográfica em cidades. Muitos sistemas de informação incluem atributos para receber endereços e, assim, contam com uma referência espacial indireta. A obtenção de coordenadas a partir de endereços é um dos métodos de geocodificação mais importantes, mas é dificultada por variações comuns no endereço, como abreviações e omissão de componentes. O artigo apresenta um método de geocodificação de endereços urbanos, que reconhece fragmentos do endereço na entrada e realiza buscas em um banco de dados geográfico de referência, para retornar coordenadas. O resultado é acompanhado de um indicador de certeza geográfica, que indica a expectativa de acerto. Uma avaliação experimental do método é apresentada.*

1. Introdução

A utilização de sistemas digitais para serviços de pesquisa, visualização de mapas, localização espacial em tempo real, está se tornando cada vez mais comum. Usuários com diversos níveis de conhecimento têm acesso fácil e rápido a esses tipos de sistemas. Esse fato traz alguns desafios para o desenvolvimento e manutenção desses sistemas, pois o ambiente, antes restrito, necessita acomodar diversos tipos de usuários com diferentes concepções sobre como realizar e buscar referências espaciais.

Dentre os diversas tipos de referências espaciais, destaca-se a realizada através de endereços postais ou urbanos. Esses endereços são compostos de fragmentos com significados diversos, como tipo do logradouro (rua, avenida, etc.), nome do logradouro,

número da edificação, bairro ou região, cidade, estado, país, código postal, etc. O uso de endereços na remessa de correspondências e na localização de pontos de interesse é rotineiro e amplamente conhecido, especialmente em cidades. Por esse motivo, endereços são usualmente incluídos como atributos em sistemas de informação convencionais. Existindo a possibilidade de obter coordenadas geográficas a partir de endereços, numa atividade conhecida como *geocodificação* (Goldberg, Wilson *et al*. 2007), tais sistemas de informação podem passar a ser geográficos.

Nem todos os sistemas de informação convencionais criam atributos diferenciados para os componentes do endereço, e é comum que o endereço seja armazenado como uma expressão textual livre (Eichelberger 1993; Davis Jr., Fonseca *et al*. 2003). Apesar da referência espacial por endereços urbanos seguir um padrão, não existem regras rígidas sobre a ordem que os componentes devem ser apresentados ou sobre elementos de separação (Rhind 1999). Isso gera dois problemas: identificação dos fragmentos de um endereço e realização de buscas a partir dos dados identificados para encontrar os resultados mais relevantes em um banco de dados de referência.

Considerando esses fatores de incerteza e possíveis causas de erros (abreviações, erros de grafia, variações de formato, entre outras), é importante que o processo de geocodificação incorpore uma medida do grau de certeza que se tem quanto ao resultado. O presente trabalho implementa e avalia um método de geocodificação de endereços urbanos proposto anteriormente (Davis Jr. and Fonseca 2007), que não apresenta uma implementação nem uma análise experimental da consistência dos resultados. O artigo está organizado da seguinte forma. A Seção 2 apresenta trabalhos relacionados, com ênfase no método de geocodificação implementado. A Seção 3 apresenta detalhes sobre a implementação e técnicas utilizadas para torná-la computacionalmente mais eficiente. A Seção 4 traz uma avaliação experimental do método. Finalmente, a Seção 5 apresenta conclusões e trabalhos futuros.

2. Trabalhos relacionados

Geocodificação é um conjunto de métodos capazes de transformar descrições em coordenadas geográficas. Essas descrições são, em geral, nomes de lugares, expressões de posicionamento relativo ou endereços, que constituem o caso mais comum. No caso de nomes de lugares, dicionários toponímicos (ou *gazetteers*) são utilizados para reconhecimento, desambiguação e localização (Hill 2000; Goodchild and Hill 2008; Machado, Alencar *et al*. 2011). Expressões de posicionamento relativo relacionam um lugar alvo a um lugar conhecido (ponto de referência), utilizando termos em linguagem natural (Delboni, Borges *et al*. 2007), como, por exemplo "hotel próximo à Praça da Liberdade, Belo Horizonte". No caso de endereços, existe uma expectativa de detalhamento hierárquico, com componentes que indicam o país, o estado, a cidade, o bairro e o logradouro, além de um código postal que sumariza esses dados. O formato de apresentação desses componentes varia de país para país, e em muitas situações, alguns componentes são intencionalmente omitidos ou simplificados.

Para contornar essa variabilidade na formação dos endereços, uma solução consiste na divisão do método em três passos ou estágios, conforme proposto por Davis e Fonseca (2007), sendo que cada estágio possui tarefas e interfaces de entrada e saída bem definidas. O primeiro estágio, chamado de *parsing*, consiste na análise léxica que leve em conta as peculiaridades da estrutura de endereços do local ou país e posterior

conversão da entrada textual contendo o endereço em uma estrutura de dados genérica. Essa estrutura de dados contém um número finito de atributos, que correspondem a cada componente do endereço. O segundo estágio, chamado de *matching*, recebe a estrutura de dados e realiza buscas em um banco de dados de referência, comparando valores por casamento exato ou aproximado de *strings* e valores numéricos, e definindo a melhor solução em caso de casamento parcial. O estágio seguinte, denominado *locating*, consiste em recuperar as referências obtidas e extrair delas as coordenadas desejadas.

Um problema na geocodificação de endereços é medir a precisão dos resultados obtidos ao fim dos três estágios. O *Geocoding Certainty Indicator* (GCI) (Davis Jr. and Fonseca 2007), representa um método para calcular a precisão e realizar a classificação dos resultados de forma a atender as necessidades do usuário do sistema. Esse índice é composto por três índices, um para cada estágio do processo de geocodificação: *Parsing Certainty Indicator* (PCI), *Matching Certainty Indicator* (MCI) e *Locating Certainty Indicator* (LCI). Em cada estágio, esses índices recebem um valor entre 0 e 1, em que 0 representa total incerteza no resultado, enquanto 1 representa máxima certeza. Esse valor é baseado em várias regras, envolvendo casamento aproximado de componentes do endereço com bonificação de acertos e desconto de erros dentre os resultados pesquisados. O GCI final é obtido através do produto dos indicadores de cada estágio.

3. Implementação da geocodificação com avaliação da qualidade

Seguindo o objetivo do presente trabalho, foi implementado o método de geocodificação proposto por Davis e Fonseca (2007), seguindo o modelo de três estágios, e utilizando o GCI para calcular o grau de certeza quanto aos resultados encontrados. As subseções a seguir descrevem detalhes sobre a implementação de cada etapa. Para maiores informações sobre o método em si, consultar o artigo original.

3.1 Estágio de *Parsing*

O estágio de *parsing* consiste em um método para identificar componentes de endereços e organizá-los em uma estrutura de dados apropriada. Para o trabalho, o método foi implementado de forma a reconhecer e estruturar entradas textuais de endereços no formato de endereço utilizado no Brasil. Esse formato possui os seguintes componentes: tipo de logradouro, nome do logradouro, número da edificação dentro de um logradouro, nome do bairro, região ou subseção de um município ou distrito, município, estado, país e código postal. Existem ainda outros atributos, tais como o nome do edifício e complementos de um endereço, porém esses atributos não estão comumente presentes ou não têm muita relevância para efeito de localização.

Para realizar o reconhecimento dos campos, o método utiliza um analisador léxico juntamente com uma análise sintática sobre os *tokens* produzidos. Essa análise procura padrões textuais que se encaixem com os campos tipo de logradouro, nome do logradouro, número da construção dentro de um logradouro e nome de região ou subseção. A análise conta com três tabelas auxiliares, que contêm um conjunto de valores usuais para tipos de logradouros, de regiões e de identificadores numéricos utilizados no endereçamento brasileiro. Além de reconhecer esses componentes, o método supõe que o restante dos *tokens* representem localizações genéricas, que podem ser bairros, municípios, estados e países, mas a interpretação desses campos é deixada a cargo do estágio de *matching*. Ao fim do processo, o *parsing* produz uma estrutura de

dados organizada contendo os componentes de endereços identificados na entrada textual.

3.2 Estágio de *Matching*

Em seguida, passamos ao estágio de *matching*, que consiste em pesquisar o valor dos campos identificados em um banco de dados de endereços a fim de realizar o melhor casamento entre os valores identificados e os dados presentes no banco de dados. O estágio foi subdividido em quatro etapas: reconhecimento de termos de localização genéricos não classificados ou não identificados no estágio de *parsing;* busca primária no banco de dados por valores que casem com campos identificados; busca complementar no banco de dados para acertar e acrescentar valores aos campos da estrutura; e aplicação de filtros numéricos sobre os resultados das etapas anteriores.

A primeira etapa do estágio de *matching* procura, dentro dos atributos de localização genérica, valores que casem com os nomes de regiões, bairros e municípios (e respectivos estados) presentes no banco de dados. Conforme o caso, os dados genéricos são transformados em nome de região ou subseção (bairros). Após o reconhecimento desses componentes, a estrutura de endereços estará completamente identificada, restando obter o casamento do nome de logradouro.

A segunda etapa consiste em realizar busca no banco de dados utilizando casamento aproximado de *strings* sobre o atributo de nome do logradouro. O algoritmo para classificar os resultados utiliza dois métodos conhecidos na literatura: *distância de Levenshtein* (ou distancia de edição) e *shift-and aproximado* (Navarro 2001). Ambos os métodos são combinados para realizar o casamento aproximado de palavras para nomes pessoais ou geográficos. Ao fim dessa etapa, um conjunto de candidatos são obtidos para prosseguir para próxima etapa. A terceira etapa recebe esses candidatos e complementa o restante dos atributos não preenchidos na busca primária com valores vindos do banco de dados. A quarta etapa consiste em determinar valor numérico mais aproximado para o número do imóvel, caso este não tenha sido localizado. Ou seja, esta etapa realiza um filtro sobre todos os números de um logradouro e escolhe aquele que possui menor distância numérica entre o valor informado e os valores existentes.

Ao longo das quatro etapas, dois indicadores que compõem o GCI são calculados. Na segunda etapa é calculado o MCI, que mede o nível de aproximação entre entrada e resultado decorrente do casamento aproximado dos *strings*. Na terceira etapa, após complementar os dados dos candidatos, é calculado o PCI, utilizando o casamento aproximado de palavras para cada campo do candidato em relação ao campo presente na estrutura de dados resultante do estágio de *parsing*.

3.3 Estágio de *Locating*

O estágio de *locating* consiste em receber os resultados do estágio *matching* e extrair coordenadas correspondentes do banco de dados de referência. Como o método apenas transforma os dados, o indicador desse estágio sempre tem valor LCI = 1 nesta implementação, e portanto o valor final do GCI é igual ao produto de PCI e MCI.

4. Avaliação experimental

Um conjunto de dados contendo entradas textuais de endereços não padronizados da cidade de Belo Horizonte foi utilizado para verificar a eficácia da implementação do

método proposto. Por entrada não padronizada entende-se entradas realizadas livremente por digitação, por parte de usuário sem qualquer conhecimento especifico de referências textuais de endereços. Foram obtidos 102 endereços textuais, todos informados em um único *string*. Em uma inspeção visual, constata-se diversos problemas, tais como erros de grafia, abreviações, ausência da indicação do tipo de logradouro e variações de formato e de sequenciamento dos componentes do endereço.

Os endereços desse conjunto foram geocodificados usando o método descrito nas seções anteriores, tendo sido obtido também o valor do GCI em cada caso. Os mesmos endereços foram fornecidos à API de geocodificação do Google Maps, e também localizados manualmente sobre o mapa da cidade, usando como referência o sistema de endereçamento pontual de Belo Horizonte. Esta última geocodificação foi adotada como *baseline* para as análises que se seguem.

Utilizamos os endereços geocodificados pelo nosso método e os comparamos com o resultado da geocodificação manual. O índice geral de acerto da geocodificação (percentual de endereços localizados corretamente pelo método) usando o método descrito foi de 85%, com GCI médio de 0,58 (desvio padrão 0,24). Usando o Google Maps, o índice de acerto foi de 66%, usando como entrada os mesmos *strings* submetidos ao nosso método. Submetemos ao Google Maps também os endereços reformatados segundo o resultado da etapa de *parsing*, e o índice de acerto aumentou para 78%, ainda abaixo do resultado obtido pelo nosso método. Na verificação manual, não foram levadas em conta eventuais erros de posicionamento geográfico dos endereços reportados pelo Google Maps, um problema analisado detalhadamente para a cidade de Belo Horizonte por Davis Jr. e Alencar (2011).

Realizamos também uma análise do valor obtido para o GCI. O objetivo foi tentar identificar um limiar a partir do qual a geocodificação tem maior confiabilidade – observando, no entanto, que aplicações diferentes podem ter exigências variáveis quanto ao nível de certeza no resultado. A Figura 1 apresenta uma comparação entre o GCI e o percentual de acerto acumulado (i.e., percentual de acerto na geocodificação de endereços com GCI menor ou igual ao valor indicado ao longo do eixo das abscissas). A curva foi obtida ordenando os endereços pelo valor de GCI correspondente, e calculando o número de acertos acumulados até aquele ponto. A forma crescente da curva indica que o GCI cumpre o seu papel, pois valores baixos de GCI correspondem a um nível menor de acerto nos resultados. A partir de GCI = 0,5, o índice de acerto já se apresenta suficientemente elevado para a maioria das aplicações; se a exigência da aplicação quanto à confiabilidade do resultado for mais alta, pode-se adotar GCI = 0,6 como limiar, e fazer verificações adicionais em endereços com GCI entre 0,4 e 0,6, descartando os endereços com GCI inferior a 0,4.

Como o GCI é formado por outros indicadores, correspondentes às etapas da geocodificação, analisamos também o comportamento do PCI e do MCI. No caso do PCI, a média obtida para este conjunto foi de 0,74, com desvio padrão de 0,20. Esse valores foram muito semelhantes aos do MCI, com média de 0,75 e desvio padrão de 0,19. Combinados com o GCI, esses parâmetros são relevantes para a análise da qualidade geral dos dados de entrada. Em conjuntos de dados mais poluídos do que o utilizado neste artigo, o PCI tenderá a ficar mais baixo, indicando a necessidade de maior padronização e controle de qualidade na entrada do dado. Por outro lado, valores baixos de MCI indicam possíveis deficiências no banco de dados de referência, ou um

acúmulo de dificuldades com nomes de logradouros ambíguos. Os valores do GCI indicam a composição desses fatores no resultado final da geocodificação.

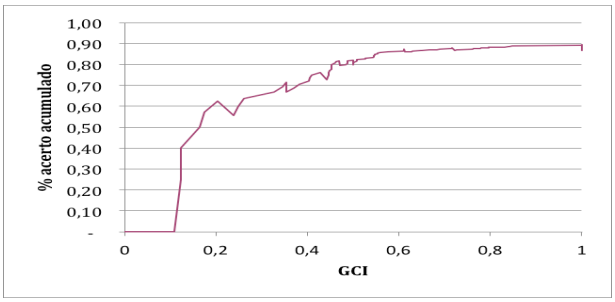

Figura 1 - GCI versus índice de acerto

5. Conclusões

O presente artigo apresentou uma implementação do método de geocodificação com verificação de confiabilidade (Davis Jr. and Fonseca 2007), acompanhada de uma verificação experimental do comportamento dos indicadores de qualidade. Os resultados foram comparados com a geocodificação oferecida na API do Google Maps, e aferidos por verificação manual. Pela análise realizada, os indicadores de qualidade da geocodificação são úteis e relevantes para as aplicações, cumprindo o papel indicado no artigo que os propôs. Trabalhos futuros envolvem a realização de avaliações mais aprofundadas, utilizando dados de entrada de qualidade variável e em maior quantidade, e a aplicação do método em situações reais, frequentemente encontradas em áreas tais como saúde pública, epidemiologia, logística e outras.

Referências

Davis Jr, C.A. and Alencar, R.O. (2011). "Evaluation of the quality of an online geocoding resource in the context of a large Brazilian city." Transactions in GIS **15**(6): 851-868.

Davis Jr., C.A., Fonseca, F. and Borges, K.A.V. (2003). A flexible addressing system for approximate urban geocoding. V Simpósio Brasileiro de GeoInformática (GeoInfo 2003), Campos do Jordão (SP):em CD-ROM.

Davis Jr., C.A. and Fonseca, F.T. (2007). "Assessing the Certainty of Locations Produced by an Address Geocoding System." Geoinformatica **11**(1): 103-129.

Delboni, T.M., Borges, K.A.V., Laender, A.H.F. and Davis Jr., C.A. (2007). "Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions." Transactions in GIS **11**(3): 377-397.

Eichelberger, P. (1993). The Importance of Addresses - The Locus of GIS. URISA 1993 Annual Conference, Atlanta, Georgia, URISA:200-211.

Goldberg, D.W., Wilson, J.P. and Knoblock, C.A. (2007). "From Text to Geographic Coordinates: The Current State of Geocoding." URISA Journal **19**(1): 33-46.

Goodchild, M.F. and Hill, L.L. (2008). "Introduction to digital gazetteer research." International Journal of Geographic Information Science **22**(10): 1039-1044.

Hill, L.L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. 4th European Conference on Research and Advanced Technology for Digital Libraries:280-290.

Machado, I.M.R., Alencar, R.O., Campos Junior, R.O. and Davis Jr, C.A. (2011). "An ontological gazetteer and its application for place name disambiguation in text." Journal of the Brazilian Computer Society **17**(4): 267-279.

Navarro, G. (2001). "A Guided Tour to Approximate String Matching." ACM Computing Surveys **33**(1): 31-88.

Rhind, G. (1999). Global Sourcebook of Address Data Management: A Guide to Address Formats and Data in 194 Countries Gower.