Uso de uma Ontologia de Lugar Urbano para Reconhecimento e Extração de Evidências Geo-espaciais na Web.

KARLA ALBUQUERQUE DE VASCONCELOS BORGES[1]

Instituição de Defesa: Departamento de Ciência da Computação da UFMG
Data da Defesa: 31/08/2006

PALAVRAS-CHAVE
Ontologia de lugar urbano - Web - busca local - endereço - evidências geo-espaciais

RESUMO

Consultas que incluem pelo menos um termo relacionado a geografia, como nomes de lugar e feições naturais, são hoje um subconjunto significativo das consultas submetidas às máquinas de busca. O interesse por informação local na Web (busca local) vem aumentando a cada dia e para esse tipo de busca, a Web é um vasto repositório de informação local e geográfica. No entanto, as máquinas de busca tradicionais apresentam limitações quanto ao reconhecimento do escopo geográfico existente nas páginas da Web. Páginas referentes ao mesmo lugar, mas que usam nomes alternativos provavelmente não serão recuperadas juntas. Além disso, muitas vezes o contexto geográfico existente nas páginas está implícito, podendo ser inferido pela existência, por exemplo, de um número de telefone ou código postal.

Para resolver esses problemas, esta tese tem como foco a *Web local*, propondo uma abordagem apoiada em uma ontologia de lugar urbano, que permita reconhecer, extrair e geocodificar evidências geo-espaciais de características locais, como endereços, códigos postais e telefones presentes em páginas da Web. As evidências geo-espaciais representam localizações implícitas, capazes de correlacionar o conteúdo de uma página, ou de parte dela, a uma localização geográfica urbana. Assim, as máquinas de busca poderiam por exemplo, utilizar essa informação para a recuperação de páginas referentes a serviços e atividades em uma determinada localidade ou próximos a ela.

Assim as principais contribuições desta tese são (1) caracterização de endereços presentes em páginas da Web como fontes de evidência geo-espacial e definição de padrões para o seu reconhecimento e extração, (2) definição da *OnLocus*, uma ontologia de lugar urbano para auxiliar o processo de reconhecimento e extração de evidências geo-espaciais de páginas da Web, (3) criação de uma base de conhecimento para reconhecimento de lugares brasileiros, baseada na *OnLocus*, (4) proposta de uma estratégia de categorização geográfica de uma página,ou de partes dela, dentro da divisão territorial de um país, e (5) avaliação das características quantitativas e qualitativas dos endereços presentes nas páginas da Web brasileira. Todas essas contribuições foram validadas por meio de experimentação, usando dados reais correspondentes a um conjunto de 4 milhões de páginas da Web. Como conseqüência adicional, foi possível traçar um retrato das páginas da Web brasileira no que tange a padrões de endereço e, conseqüentemente, entender melhor como geocodificá-las. Os resultados desta tese abrem um leque de perspectivas para novos tipos de aplicação, como, por exemplo, uso de *links* de navegação baseados em localização geográfica, classificação geográfica das páginas da Web, mineração de dados geo-espaciais em páginas da Web e anotação semântica das páginas.

[1] E-mail: karla@pbh.gov.br

Keywords
Ontology of urban places - Web - local search - geoespatial evidences

Abstract

Queries that include at least one geographic-related term, such as place names and natural features, are currently a significant subset of the queries that are submitted to search engines. Interest on local information on the Web (local search) is increasing daily, and for this kind of search, the Web is a vast repository of local geographic information. However, traditional search engines have limitation on the recognition of the geographic scope of Web pages. Pages that refer to the same place, but using alternative names, probably will not be retrieved together. Besides, in many situations the geographic context is implicit in the pages, but can be inferred by the existence, for instance, of a telephone number or postal code.

In order to propose a solution for these problems, this thesis focuses on the *local Web*, presenting an approach based on an ontology of urban place, which allows for the recognition, extraction, and geocoding of geospatial evidences with local characteristics, such as urban addresses, postal codes, and telephone numbers as found in Web pages. The geospatial evidences are implicitly related to places, so that the contents of a page, or parts of it, can be correlated to an urban geographic location. Thus, search engines can, for instance, use such information to retrieve pages that are related to services and activities in a certain location or close to it.

Therefore, the main contributions of this thesis are (1) the characterization of urban addresses contained in Web pages as sources of geospatial evidences and definition of patterns for their recognition and extraction, (2) the definition of *OnLocus*, an ontology of urban place that helps in the process recognizing and extracting geospatial evidences from Web pages, (3) the creation of a database for recognition of Brazilian places, based on *OnLocus*, (4) the proposal of a strategy for geographic categorization of a Web page, or parts of it, within a country's territorial divisions, and (5) the evaluation of the quantitative and qualitative characteristics of urban addresses that are found in the pages of the Brazilian Web. All of these contributions have been validated through experimentation, using real data from a set of 4 million Web pages. As an additional result, it was possible to obtain a snapshot of the usage of addresses in pages from the Brazilian Web and, consequently, to better understand how to geocode them. Results of this thesis open a range of perspectives for new types of applications, such as, for instance, the use of navigational links based on geographic location, geographic classification of Web pages, Web-based geospatial data mining, and semantic annotation of pages.

Sobre a autora

Karla Albuquerque de Vasconcelos Borges
Graduada em Engenharia Civil pela Pontifícia Universidade Católica – 1982
Mestre em Administração Pública com ênfase em Informática pela Escola de Governo – Fundação João Pinheiro – 1997
Doutora em Ciência da Computação pela Universidade Federal de Minas Gerais – 2006
Analista de Informática da Empresa de Informática e Informação do Município de Belo Horizonte - PRODABEL
karla@pbh.gov.br