Expansão do conteúdo de um *gazetteer*: nomes hidrográficos

Tiago Henrique V. M. Moura, Clodoveu A. Davis Jr

Departamento de Ciência da Computação - Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte, MG - Brasil

[thvmm,clodoveu]@dcc.ufmg.br

Abstract. *The efficiency of a geographic database is directly related with the quality and completeness of its contents. In the case of gazetteers, i.e., place name dictionaries, previous work proposed ontological extensions based on the storage of geographic shape and on multiple types of relationships among places. However, in order to be more useful, gazetteers must contain large volumes of information on a large variety of themes, all of which must be geographically represented and related to places. The objective of this work is to propose techniques to expand a gazetteer's content using relevance criteria, increasing its usefulness to solve problems such as place name disambiguation. We demonstrate these techniques using data on Brazilian rivers, which are preprocessed, and the appropriate relationships are identified and created.*

Resumo. *A eficiência de um banco de dados geográficos está diretamente relacionada à qualidade e completude das informações nele contidas. No caso de gazetteers, i.e., dicionários de nomes de lugares, trabalhos anteriores propuseram extensões ontológicas baseadas no armazenamento das formas geométricas e na criação de múltiplos tipos de relacionamentos entre lugares. No entanto, para que tenham maior utilidade, os gazetteers precisam conter grandes volumes de informação sobre uma variedade de temas relacionados a lugares. O objetivo deste trabalho é propor técnicas para expandir o conteúdo de um gazetteer usando critérios de relevância, aumentando sua utilidade em problemas como a desambiguação de nomes de lugares. É apresentado um estudo de caso com dados de rios brasileiros, que são pré-processados e incluídos no gazetteer, juntamente com os relacionamentos apropriados.*

1. Introdução

O volume de informação disponível na internet atualmente é muito grande e cresce diariamente. Buscar tal informação requer sistemas capazes de compreender o que o usuário deseja, localizar e apresentar resultados em ordem de relevância. Muitas vezes o usuário utiliza um conjunto de palavras-chave como forma de dizer o que procura para o sistema. Trabalhos anteriores (Sanderson and Kohler 2004; Wang, Xie *et al*. 2005; Delboni, Borges *et al*. 2007; Backstrom, Kleinberg *et al*. 2008) mostram que uma parte significativa dessas consultas envolve termos como nomes de lugares e expressões que denotam posicionamento. Por isso, é importante reconhecer a intenção do usuário que inclui termos geográficos em buscas, bem como determinar o escopo geográfico de documentos, em aplicações de recuperação de informação geográfica (RIG).

Em problemas de RIG, é frequentemente necessário reconhecer um nome como sendo uma referência a um lugar, e também distinguir entre lugares que possuem o

mesmo nome (Hastings 2008). Por exemplo, "São Francisco" pode ser uma cidade da região norte de Minas Gerais, um bairro de Belo Horizonte, um rio ou um santo católico. Os *gazetteers* (dicionários toponímicos) são recursos que auxiliam nesse processo. Visando RIG e outras aplicações, nosso grupo projetou e desenvolveu um *gazetteer* ontológico, denominado *Ontogazetteer* (Machado, Alencar *et al*. 2011), em que não apenas são registrados nomes de lugares, mas também relacionamentos entre eles. Nesse *gazetteer*[1] estão também incluídos dados urbanos, utilizados cotidianamente pelos cidadãos, particularmente em mensagens disseminadas nas redes sociais online.

Este trabalho apresenta técnicas para expandir o conteúdo de um *gazetteer* usando critérios de relevância, voltadas especificamente para o Ontogazetteer. Trabalhos relacionados são descritos na Seção 2. Um estudo de caso envolvendo rios brasileiros é apresentado na Seção 3, sendo definidos também os relacionamentos apropriados (Seção 4). Finalmente, a Seção 5 apresenta conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Em geral, *gazetteers* contêm dados organizados segundo uma tripla `<nome do lugar, tipo do lugar, footprint>`, sendo que esse *footprint*, que representa a localização geográfica propriamente dita, se resume a um par de coordenadas (Hill 2000). Exemplos de *gazetteers* com essa estrutura básica incluem o GeoNames e o Getty Thesaurus of Geographical Names (TGN). Tais *gazetteers* são utilizados como fontes de nomes geográficos para diversas aplicações (Souza, Davis Jr. *et al*. 2005; Goodchild and Hill 2008). A principal função desses *gazetteers* é informar uma coordenada geográfica a partir de um nome de lugar dado, o que os torna apenas parcialmente adequados às necessidades de RIG.

O Ontogazetteer (Machado, Alencar *et al*. 2010; Machado, Alencar et al. 2011) foi proposto com uma estrutura mais complexa que a usual, em que os lugares (1) podem ser representados por pontos, linhas ou polígonos, (2) estão relacionados a outros lugares, usando relacionamentos espaciais (vizinho a, contido em, etc.) ou semânticos, (3) podem possuir nomes alternativos ou apelidos, e (4) podem estar associados a termos e expressões características (Alencar and Davis Jr 2011). Tais características adicionais são importantes para RIG, pois fornecem elementos para resolver problemas importantes, como a ambiguidade de nomes de lugares (Leidner 2007) e a detecção do contexto geográfico em textos (Silva, Martins *et al*. 2006). A expansão do conteúdo desse modelo semanticamente mais rico de *gazetteer* é um desafio importante, para ampliar a gama de situações em que técnicas de RIG poderão ser usadas para reconhecimento de lugares associados a textos. Para a expansão, podem ser utilizados dados extraídos de bancos de dados geográficos, filtrando o que é irrelevante, e detectando relacionamentos com lugares anteriormente disponíveis. A decisão quanto ao que é ou não relevante para ser incorporado ao *gazetteer* precisa levar em conta critérios baseados nas características dos lugares ou de seus relacionamentos com outros lugares. Diante do exposto, este trabalho busca expandir o conteúdo do OntoGazetteer, não apenas acrescentando novos nomes de lugares, mas também aumentando e diversificando os relacionamentos entre esses lugares.

[1] http://geo.lbd.dcc.ufmg.br:8080/ontogazetteer/

3. Expansão

Uma fonte para lugares e relacionamentos são bancos de dados geográficos existentes, dos quais se pode extrair nomes relacionados a objetos geográficos e determinar relacionamentos com base em sua geometria. A obtenção de relacionamentos semânticos, por outro lado, é mais complexa, pois sua natureza pode ser muito variada. Por exemplo, as cidades de Perdões (MG) e Bragança Paulista (SP) estão relacionadas por serem cortadas pela BR-381, embora não sejam vizinhas nem se localizem próximas uma a outra. Da mesma forma, lugares geograficamente desconexos podem ter ligações semânticas baseadas em características em comum (p. ex. estâncias hidrominerais de Caxambu (MG) e Poá (SP)), ou formarem grupos semanticamente coerentes (p. ex., Pico da Neblina e Pico da Bandeira), ou ainda por motivos históricos (p. ex., Mariana, Ouro Preto e Belo Horizonte, capitais de Minas Gerais ao longo da história).

Relacionamentos espaciais semânticos constituem uma vantagem para o uso do Ontogazetteer em diversas aplicações (Machado, Alencar et al. 2011). Por exemplo, uma notícia que contenha os nomes "Salvador", "Camaçari" e "Dias D'Ávila" provavelmente se refere à Região Metropolitana de Salvador, unidade espacial de referência que contém municípios com esses nomes. Outra notícia que contenha os nomes "Sabará", "Cordisburgo" e "Curvelo", mesmo que esses nomes sejam devidamente associados aos municípios correspondentes, teria seu escopo geográfico definido como "Minas Gerais", referência que contém todos os três municípios. Estando registrado um relacionamento semântico baseado em rios e bacias hidrográficas, por outro lado, seria possível concluir, com mais precisão, que o escopo na verdade é a bacia do rio das Velhas, afluente do rio São Francisco que passa pelos três municípios. O rio das Velhas, no caso, constitui uma conexão semântica entre as três cidades.

Assim, este trabalho introduz técnicas para a expansão do conteúdo do OntoGazetteer, com foco particular sobre relacionamentos semânticos. Um primeiro estudo foi realizado sobre dados de rios e bacias hidrográficas do Brasil publicados pela Agência Nacional de Águas (ANA) em seu Web site e busca obter relacionamentos semânticos entre lugares que estejam direta ou indiretamente relacionados a rios e bacias hidrográficas. Os elementos, rios e bacias, existentes nesta base foram codificados seguindo a proposta de Otto Pfafstetter (ANA 2006) e obedecem uma hierarquia onde nos níveis mais altos estão os rios que deságuam no oceano. Esses dados precisam ser transformados para carga no *gazetteer*, pois apresentam alguns problemas, como a falta dos nomes de alguns elementos. Por isso, uma série de filtros foram aplicados a fim de se obter os rios e bacias mais relevantes.

O primeiro filtro executado removeu elementos com nomes indeterminados, reduzindo o volume de dados em mais de 50%. Entretanto, em análises mais detalhadas constatou-se que este filtro precisava ser revisto, devido à existência de rios que passam por regiões densamente habitadas e estavam sem nome na base da ANA. Um exemplo é o Ribeirão Arrudas, que cruza a cidade de Belo Horizonte, e que tem pequena importância hidrológica por ter pequeno comprimento e baixa vazão, mas importante devido à intensa urbanização em sua bacia, que o transforma em uma referência urbana.

Outros cursos d'água se encontravam na mesma situação. Para encontrar essas situações, e buscar resolvê-las com dados de outras fontes, buscou-se estabelecer o valor de um elemento com nome indeterminado. No caso, optou-se por considerar como importantes rios que, mesmo sem nome definido e de pequeno porte, cruzassem

municípios cuja população total excedesse 3 milhões de habitantes. Um total de 49 rios atenderam a tal critério e, utilizando ferramentas auxiliares como Google Maps, Wikipedia e Wikimapia, 18 nomes foram determinados e utilizados. Para assegurar a corretude dessa ação, foram considerados rios afluentes, pontos de deságue, localização geográfica e municípios vizinhos ao elemento. Destes critérios, o que mais trouxe resultados foi a relação dos rios com outros rios que têm seus nomes no banco, como por exemplo vários afluentes de menor porte do rio Tietê. Outro critério bem sucedido foi o relacionamento topológico com os municípios que são interceptados pelo rio.

Outro problema existente nos dados da ANA era a forma segundo a qual os rios estavam hierarquizados. A hierarquia a qual os dados obedecem para fins de codificação não era condizente com a relevância dos dados para o dicionário geográfico. Os rios foram classificados em sete níveis, sendo o nível mais alto (nível 1) o rio que deságua no mar e o mais baixo o afluente mais distante do mar (ANA 2006). Foi, então, proposta uma nova hierarquização que permitisse selecionar rios de maior importância do ponto de vista do reconhecimento de seu nome. Inicialmente, essa reclassificação baseou-se apenas em dados geográficos, como comprimento do rio ou área de sua bacia. Após está primeira tentativa de classificação, constatamos que níveis inferiores continham um grande número de rios e dentre eles existia uma diferenciação de relevância; por exemplo, um pequeno rio que corta a capital de um estado é mais importante para o *gazetteer* que um grande igarapé na floresta amazônica. Para resolver essa questão, assim como na classificação de elementos sem nomes, foram utilizados dados demográficos do IBGE juntamente com filtros que consideram apenas dados geográficos. A Tabela 1 mostra duas regras distintas utilizadas para filtrar e reclassificar os elementos existentes na base da ANA, onde A é a área da bacia em Km², C o comprimento do rio em Km e P a população atendida pelo rio.

Tabela 1. Filtros implementados

	Regra Baseada na Área da Bacia (A) em Km²	Regra Baseada no Comprimento do Rio (C) em Km
Nível 1	$A > 100000$	$C > 1150$
Nível 2	$10.000 < A \leq 100.000$	$550 < C \leq 1150$
Nível 3	$(2.000 < A \leq 10.000)$ && $(P \geq 50.000)$	$(150 < C \leq 550)$ && $(P \geq 50.000)$
Nível 4	$(1.000 < A \leq 2.000)$ && $(P \geq 50.000)$	$(0 < C \leq 150)$ && $(P \geq 50.000)$
Nível 5	$(0 < A \leq 1.000)$ && $(P \geq 50.000)$	–

A separação obedecendo à área da bacia obteve melhores resultados comparada ao critério baseado no comprimento do rio, pois dentre os níveis criados pode-se notar uma melhor padronização nas características dos rios. A utilização de critérios demográficos só foi necessária nos níveis inferiores ao segundo nível.

As bacias hidrográficas também foram incorporadas ao *gazetteer*. Para isso, foram associadas ao nome de seu principal rio. Como muitos rios foram desconsiderados para o *gazetteer*, também apenas as bacias relevantes e com nome significativo foram incorporadas. Dos 178.561 trechos de rios e 77.859 bacias disponíveis nos dados da ANA, foram incorporados ao gazetteer um total de 5.384 rios e 670 bacias. O resultado final do processo de filtragem demonstra a redução significativa do número de elementos considerados, sem perda de dados relevantes para o *gazetteer*, uma vez que foram preservados todos os nomes geográficos encontrados e acrescentados alguns outros.

4. Relacionamentos

A parte que mais agrega valor ao *gazetteer* é a criação dos relacionamentos entre as entidades existentes no mesmo. Por isso, a tarefa de estipular quais seriam criados foi feita cuidadosamente. Foi definido foi que a menor unidade espacial com a qual um rio ou bacia deveria se relacionar seria um município.

Foram definidos 18 novos tipos de relacionamentos para o *gazetteer* envolvendo rios e bacias, divididos em três grupos: o primeiro relaciona espacialmente rios e bacias correspondentes, o segundo relaciona espacialmente rios e bacias com os demais elementos do *gazetteer* e o terceiro relaciona semanticamente os elementos do *gazetteer* que estão relacionados por intermédio de rios e/ou bacias comuns entre eles. A Tabela 2 lista os 18 novos tipos de relacionamentos criados.

Tabela 2. Relacionamentos criados

Ent1	Relacionamento	Ent2	Gr
Rio	Afluente de	Rio	1
Bacia	Contida em	Bacia	1
Rio	Parte de	Bacia	1
Rio	Intercepta	Estado	2
Rio	Intercepta	Município	2
Município	Intercepta	Bacia	2
Estado	Intercepta	Bacia	2
Rio	Intercepta	Mesorregião	2
Rio	Intercepta	Microrregião	2

Ent1	Relacionamento	Ent2	Gr
Rio	Intercepta	Macrorregião	2
Mesorregião	Intercepta	Bacia	2
Microrregião	Intercepta	Bacia	2
Macrorregião	Intercepta	Bacia	2
Mesorregião	Int. pelo mesmo rio	Mesorregião	3
Microrregião	Int. pelo mesmo rio	Microrregião	3
Macrorregião	Int. pelo mesmo rio	Macrorregião	3
Município	Int. pelo mesmo rio	Município	3
Estado	Int. pelo mesmo rio	Estado	3

Com esses relacionamentos, o grafo de relacionamento entre as entidades é expandido para envolver boa parte do que já existe atualmente no *gazetteer*, aumentando o potencial da ferramenta na solução de problemas. Naturalmente, na medida em que novos tipos de entidades vão sendo incorporados ao *gazetteer*, a construção de relacionamentos fica mais complexa, simplesmente pelo efeito de combinação das entidades duas a duas. No entanto, a existência da definição do tipo de relacionamento permite às aplicações considerar apenas parte dos relacionamentos.

5. Conclusões e Trabalhos Futuros

Este artigo descreveu as etapas realizadas no processo de expansão de um *gazetteer* partindo de dados da Agência Nacional de Águas (ANA) sobre rios e bacias hidrográficas. A partir da forma na qual os dados foram originalmente organizados foram aplicados sucessivos filtros para se obter o subconjunto de elementos que agregassem mais valor à capacidade de resolução de problemas do *gazetteer*. Na construção dos filtros ficou evidente a necessidade de utilizar dados auxiliares para a determinação da importância dos elementos. Foram utilizados dados demográficos e as próprias informações da ANA, como comprimento de rios e área de bacias.

Um registro no *gazetteer* só faz sentido se este está relacionado a um nome de lugar real e reconhecível pelas pessoas. Por isso, pretende-se realizar no futuro uma análise mais detalhada dos rios que estavam sem nome nos dados da ANA e cujos nomes não pudemos identificar. Uma alternativa é usar contribuições voluntárias (Silva and Davis Jr 2008; Twaroch and Jones 2010), de modo que cidadãos com conhecimento local possam ajudar nessa determinação. Para se obter um resultado ainda melhor seria necessária a expansão também de outras relações do *gazetteer*, que guardam informações como nomes ambíguos, termos relacionados e nomes alternativos.

Destacamos que as técnicas apresentadas aqui estão sendo utilizadas em outras expansões, envolvendo elementos tais como rodovias, ferrovias e lugares agrupados segundo categorias encontradas em bases de conhecimento tais como a Wikipedia.

Agradecimentos

Este trabalho foi parcialmente financiado com recursos do CNPq (302090/2009-6 e 560027/2010-9) e FAPEMIG (CEX-PPM-00466/11), além do Instituto Nacional de Ciência e Tecnologia para a Web (InWeb, CNPq 573871/2008-6).

Referências

Alencar, R.O. and Davis Jr, C.A. (2011). Geotagging aided by topic detection with Wikipedia. 14th AGILE Conference on Geographic Information Science, Utrecht, The Netherlands:461-478.

ANA (2006). Topologia hídrica: método de construção e modelagem da base hidrográfica para suporte à gestão de recursos hídricos. Agência Nacional de Águas. Brasília (DF). **Versão 1.11, 17/11/2006**.

Backstrom, L., Kleinberg, J., Kumar, R. and Novak, J. (2008). Spatial Variation in Search Engine Queries. International World Wide Web Conference (WWW), Beijing, China:357-366.

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.

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

Hastings, J.T. (2008). "Automated conflation of digital gazetteer data." International Journal of Geographical Information Science **22**(10): 1109-1127.

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.

Leidner, J.L. (2007). Toponym Resolution in Text: annotation, evaluation and applications of spatial grounding of place names. Boca Raton, Florida, Dissertation. com.

Machado, I.M.R., Alencar, R.O., Campos Junior, R.O. and Davis Jr, C.A. (2010). An Ontological Gazetteer for Geographic Information Retrieval. XI Brazilian Symposium on Geoinformatics, Campos do Jordão (SP), Brazil:21-32.

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.

Sanderson, M. and Kohler, J. (2004). Analyzing Geographic Queries. Proc. of the ACM SIGIR Workshop on Geographic Information Retrieval, Sheffield, UK:1-2.

Silva, J.C.T. and Davis Jr, C.A. (2008). Um framework para coleta e filtragem de dados geográficos fornecidos voluntariamente. X Brazilian Symposium on GeoInformatics (GeoInfo 2008), Rio de Janeiro (RJ), Sociedade Brasileira de Computação.

Silva, M.J., Martins, B., Chaves, M., Cardoso, N. and Afonso, A.P. (2006). "Adding Geographic Scopes to Web Resources." Computers, Environment and Urban Syst. **30**: 378-399.

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

Twaroch, F.A. and Jones, C.B. (2010). A Web Platform for the Evaluation of Vernacular Place Names in Automatically Constructed Gazetteers. 6th International Workshop on Geographical Information Retrieval (GIR 2010), Zurich, Switzerland.

Wang, C., Xie, X., Wang, L., Lu, Y. and Ma, W. (2005). Detecting Geographic Locations from Web Resources. Proc. of the 2nd Int'l Workshop on Geographic Information Retrieval, Bremen, Germany:17-24.