Desambiguação de Topônimos Usando Dicionários Geográficos

Celina Maki Takemura; Gustavo Bayma-Silva; Stanley Robson de Medeiros Oliveira; Maria Fernanda Moura

Resumo

O objetivo deste trabalho foi apresentar uma metodologia de extração e desambiguação de topônimos a fim de georreferenciar publicações técnico-científicas através da identificação de regiões geográficas, tais como: bacias hidrográficas, corpos d'água, estados, municípios e biomas associados. A metodologia prevê o (1) reconhecimento de entidades nomeadas; (2) inferência sobre relação das entidades nomeadas com os termos de um *gazetteer*, i.e., índice de topônimos; e (3) um processo de desambiguação baseado em distâncias.

Termos para indexação: informação geoespacial, mineração de textos, inferência espacial.

Toponym Disambiguation Using Gazetteers

Abstract

In this work, we present a methodology for extraction and disambiguation of toponyms in order to georeference scientific publications through the identification of watersheds, water bodies, states, municipalities and biomes. The methodology provides for (1) named entity recognition; (2) inference about the relationship between the named entities and the terms of a gazetteer, i.e., toponyms index; and (3) a disambiguation process based on distances.

Index terms: geospatial information, text mining, spatial inference

Introdução

O alinhamento de conceitos geoespaciais (i.e., localização: cidade, estado, país), tipos de lugares (bioma, região hidrográfica, município) à informação geoespacial (e.g., polígono envolvente) é importante e necessário em atividades de planejamento, gestão de recursos, tomada de decisão e para a elaboração de políticas públicas. A extração da geoinformação em textos tem importância estratégica para inúmeros setores, incluindo o fortalecimento da agricultura brasileira e a análise dos impactos do uso agrícola e das mudanças climáticas sobre os recursos hídricos em diferentes ecorregiões brasileiras.

Neste contexto, o trabalho desenvolvido junto à Rede AgroHidro para a extração, identificação e desambiguação de topônimos em publicações científicas utilizadas e produzidas pela Rede, visa a georreferenciação de textos para diversos fins, tais como: (1) a análise da distribuição espacial das publicações em determinados assuntos (e.g, "aumento das temperaturas" ou "manejo de água e solo", etc); (2) auxílio à construção de um portfólio de tecnologias que indiquem soluções tecnológicas para o uso sustentável da água na agricultura, levando em consideração particularidades do manejo em diferentes escalas (e.g., bacia hidrográfica ou região; perímetro irrigado; parcela). Dessa forma, a metodologia apresentada procura associar a textos polígonos delimitadores de regiões geográficas.

Material e Métodos

Os documentos utilizados e produzidos pela Rede AgroHidro são, em sua maioria, disponíveis em provedores de dados de livre acesso no padrão OAI[1]. Foram selecionados dentre esses provedores os repositórios de publicações técnico-científicas da Embrapa e de diversos jornais e revistas do domínio agrícola de maior interesse para a Rede. Como entrada da metodologia apresentada nesse resumo, foram utilizados arquivos recuperados dos provedores selecionados.

[1] Protocolo Open Archives Initiative (Iniciativa de Arquivos Abertos)/Dublin Core (DCMI, 2016).

Para georreferenciar o arquivo, são utilizados os campos OAI/DCMI: título, descrição (abstract/resumo) e, quando disponível, o texto completo da publicação (normalizado para a codificação UTF-8). A partir deles, as entidades nomeadas são extraídas com o uso de uma ferramenta linguística. Essas entidades nomeadas podem ser locais, nomes próprios, datas, tecnologias, termos da indústria, etc. A ferramenta OpenCalais (THOMSON REUTERS, 2016) foi utilizada como Reconhecedor de Entidades Nomeadas (REN). OpenCalais trabalha com textos em inglês, portanto, faz-se necessário um processo de tradução para os textos em outras línguas e, neste trabalho, utilizamos o pacote de processamento de língua natural TEXTBLOB (LORIA, 2016) para Python.

Encontradas as entidades nomeadas, o interesse é desambiguar aquelas reconhecidas como '*NaturalFeature*', '*ProvinceOrState*' e '*City*'. Por exemplo, é necessário identificar se a cidade Bom Jesus, citada em um texto, é a Bom Jesus da Paraíba, do Piauí, do Rio Grande do Norte, de Santa Catarina ou do Rio Grande do Sul. Para tanto, calcula-se a similaridade de cada entidade com cada um dos termos do índice de topônimos. Foram usados para a geração do índice de topônimos, com 19.013 termos, os arquivos vetoriais em formato *shapefile*: (a) Malha municipal brasileira contendo feições poligonais da divisão político-administrativa brasileira (IBGE, 2007); (b) Malha estadual brasileira, derivada de (IBGE, 2007); (c) Mapa Temático de Biomas do Brasil (IBGE, 2004); (d) Rede hidrográfica brasileira, contendo feições lineares de recursos hídricos topologicamente tratadas e codificadas pelo método de Otto Pfafstetter (AGÊNCIA NACIONAL DE ÁGUAS, 2013); (e) Mapas das Grandes Bacias Hidrográficas brasileiras (AGÊNCIA NACIONAL DE ÁGUAS, 2016).

Para normalizar a representação da região geográfica, cada topônimo é associado a um polígono ou área de interesse no sistema de referência SAD69, i.e., o envelope da geometria que o representa, gerando um arquivo com o tipo (referência ao mapa), um identificador único, o topônimo em si e o polígono, usado no processo de desambiguação.

Ou seja, para cada documento extrai-se um conjunto de n entidades, $E=\{e_1,..,e_n\}$, e para cada, e_i, i=1,…,n, um conjunto associado de polígo-

nos $P_i=\{p_{i1},\ldots,p_{iti}\}$, j=1,..,$t_i$, onde t_i é o número de possíveis topônimos correspondentes à i-ésima entidade, e_i. O processo de desambiguação consiste em encontrar o polígono envolvente, entre os possíveis topônimos, que minimiza a distância entre todos os possíveis polígonos das diferentes entidades. Para isso, gera-se um arranjo T de todos os possíveis polígonos p_{ij}, i=1,..n, e j=1,..t_i com i≠j. Assim tem-se $T=\{G_1,\ldots,G_N\}$, onde N é o número de possíveis polígonos envolventes, com $G_k=\{p_1,..p_K\}$ e calcula-se as somas das distâncias euclidianas entre cada $p_1,..,p_K$ em G_k; lembrando que cada p_k está em um P_i, i=1,..n. O conjunto de polígonos desambiguados será o G_k com a menor soma de distâncias.

Resultados e Discussão

Seja o documento abaixo, com o título e a descrição (resumo/abstract):

> <Title> Sistemas de produção praticados e sistemas melhorados propostos - Bacia do Rio Formoso, Bonito, MS. </Title>
> <Description> O presente trabalho faz parte do Projeto "Gestão Integrada da Bacia Hidrográfica do Rio Formoso", financiado pelo Fundo para o Meio Ambiente Mundial (Global Environment Facility - GEF) e implementado pelo Banco Internacional para a Reconstrução e o Desenvolvimento (Bird). O objetivo deste trabalho foi de caracterizar os sistemas de produção modais (predominantes) praticados pelos produtores rurais da Bacia do Rio Formoso, Bonito, MS e a partir destes, propor sistemas melhorados, mais sustentáveis. As informações para a caracterização de cada sistema foram levantadas por meio de painéis participativos, com a presença de técnicos, pesquisadores, produtores rurais e outros interessados. Foram realizados em 2006, quatro painéis para caracterização dos sistemas... </Description>

O reconhecedor de entidades nomeadas retorna, para a versão traduzida do texto:

> entities (NaturalFeature): Formoso River
> entities (City): Bonito

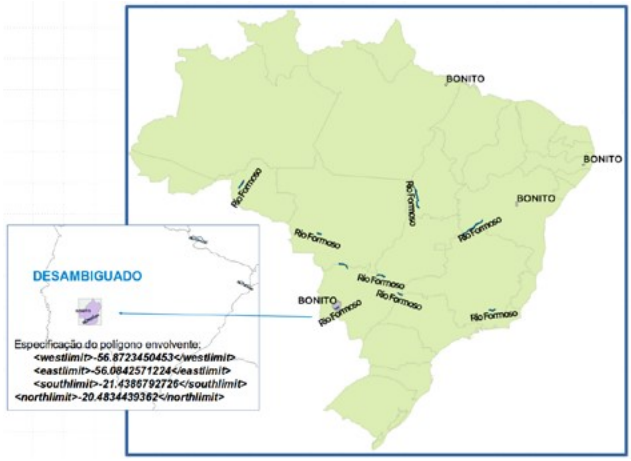

Figura 1. Resultado do processo de extração de topônimos (Rio Formoso e Município de Bonito selecionados e o envelope envolvente resultante).

O resultado do processo de desambiguação, expresso na forma do envelope envolvente das estruturas e dos termos selecionados, é mostrado na Figura 1. Nota-se, neste exemplo, que os erros acumulados na tradução para inglês e o próprio processamento do reconhecedor de entidades nomeadas pode trazer prejuízos ao resultado da extração e associação de topônimos, bem como da geometria da área de interesse de cada documento. Note que o texto cita o Estado do Mato Grosso do Sul (representado na sigla MS) e Bacia do Rio Formoso, que não aparecem nas entidades nomeadas e consequentemente não são utilizadas no processo de desambiguação e, por fim, não constam do resultado. Todavia, observa-se que o resultado, para um processo sem supervisão, é subjetivamente julgado satisfatório.

Conclusões

Esse resumo apresenta a metodologia utilizada para georreferenciar textos de interesse da Rede AgroHidro. O processo implementado a partir desta metodologia pode auxiliar a análise de grandes volumes de documentos combinando extração de topônimos e representação geo-

espacial dos mesmos. A automatização do processo implica em alguns erros, porém a maior parte dos resultados encontrados, aleatoriamente selecionados para conferência, tem sido satisfatória. Espera-se, para trabalhos futuros, a inclusão de outros shapefiles no índice de topônimos, com informações tais como meso e microrregiões e sub-bacias, para obter da melhor forma a escala de manejo de solo e água a que os textos se referem. Em adição, deverão ser avaliados outros pacotes para a tradução automática dos textos, visando aumentar a eficácia do reconhecedor de entidades nomeadas utilizado, bem como testar alguns reconhecedores de entidades nomeadas para a língua portuguesa. Por fim, devem ser realizados experimentos de validação junto a especialistas de domínio.

Referências

AGÊNCIA NACIONAL DE ÁGUAS. **Bacias hidrográficas brasileiras**. Disponível em: <http://hidroweb.ana.gov.br/HidroWeb.asp?TocItem=4100>. Acesso em: 9 maio 2016.

AGÊNCIA NACIONAL DE ÁGUAS. **Rede hidrográfica brasileira**. Disponível em: <http://www.ana.gov.br/bibliotecavirtual/redeHidrografica.asp>. Acesso em: 9 set. 2013.

DUBLIN CORE METADATA INITIATIVE. **CDMI Home:** Dublin Core Metadata Initiative **(**DCMI**)**. Disponível em: <dublincore.org>. Acesso em: 9 maio 2016.

EMBRAPA. **SABIIA – Sistema Aberto e Integrado de Informação em Agricultura**. Disponível em: <https://www.sabiia.cnptia.embrapa.br/>. Acesso em: 9 maio 2016.

IBGE. **Malha municipal brasileira**. Disponível em: <ftp://geoftp.ibge.gov.br/malhas_digitais/municipio_2007/escala_2500mil/ proj_geografica_sad69/brasil/55mu2500gsd.zip>. Acesso em: 9 set 2013.

IBGE. **Mapa temático**: mapa de biomas do Brasil. Disponível em: <ftp://geoftp.ibge.gov.br/mapas_tematicos/mapas_murais/shapes/biomas>. Acesso em: 9 maio 2016.

LORIA, S.; KEEN, P., HONNIBAL, M.; YANKOVSKY, R.; KARESH, D.,; DEMPSEY, E.; CHILDS, W.; SCHNURR. J.; QALIEH, A.; RAGNARSSON, L.; COE, J.; CALVO, A.L. **TextBlob**: simplified text processing. [Charlottesville, 2016]. Disponível em: < https://textblob.readthedocs.io>. Acesso em: 9 maio 2016.

MOURA, M. F.; TARARAM, G. M.; MARCACINI, R. M.; GONZALES, L. E.; TAKEMURA, C. M.; SILVA, L. E. A.; SANTOS, F. F. dos; REZENDE, S. O.; EVANGELISTA, S. R. M. Um software para recuperar e analisar artigos Open Access em agricultura utilizando técnicas de mineração de textos. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 10., 2015, Ponta Grossa. **Uso de VANTs e sensores para avanços no agronegócio**: anais... Ponta Grossa: UEPG, 2015.

THOMSON REUTERS. **OpenCalais**. Disponível em: <http://www.opencalais.com/>. Acesso em: 9 maio 2016.

VACARI, I.; VISOLI, M. C.; GONZALES, L. E. Acesso aberto a informação científica agropecuária na internet: caso do sistema aberto e integrado de informação em agricultura (Sabiia). In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 8., 2011, Bento Gonçalves. **Anais...** Florianópolis: UFSC; Pelotas: UFPel, 2011.