Modelagem e desenvolvimento de reconhecedor geoespacial para documentos textuais

Lucas Siegl Correa Machado[1]
Maria Fernanda Moura[2]
Celina Maki Takemura[3]

Este trabalho tem como objetivo modelar e desenvolver um software para a geoespacialização dos documentos relacionados ao projeto Compilação e Recuperação de Informações Técnico-científicas e Indução ao Conhecimento de forma Ágil na Rede Agrohidro (CRITIC@), agregando-lhes metadados geoespaciais. Desta forma será possível a realização de análises, tais como observar as relações entre as tecnologias utilizadas e a disponibilidade hídrica em uma dada região.

Para identificar e desambiguar topônimos nos textos dos documentos, foi adaptado o método Spatial Coverage Identification Methodology (SpatialCIM) (VARGAS et al., 2012). Apesar dos bons resultados obtidos com as implementações da SpatialCIM a adaptação foi criada, afim de suprir algumas limitações existentes no método. Nele a identificação de candidatos a topônimos nos textos estava restrita a uma ferramenta que trabalha com a língua portuguesa, a Rembrandt (CARDOSO, 2008), ainda, o desambiguador de topônimos implementado para a SpatialCIM utiliza-se de um único ponto (longitude/latitude) para geolocalização do topônimo e de um gazetteer restrito aos nomes geográficos das divisão político administrativa do Brasil utilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

Essa adaptação prevê a expansão para a língua inglesa, a utilização de nomes geográficos relacionados à hidrografia do Brasil e a exten-

[1] Universidade Estadual de Campinas - lucas.machado@colaborador.embrapa.br
[2] Embrapa Informática Agropecuária - maria-fernanda.moura@embrapa.br
[3] Embrapa Monitoramento por Satélite - celina.takemura@embrapa.br

são do tratamento de pontos para polígonos delimitadores (geometrias). Para identificação dos topônimos, foram testados os seguintes reconhecedores de entidades nomeadas para o inglês: OpenCalais (THOMPSON REUTERS, 2008), NLTK (DAN GARRETE et al., 2005) e Stanford NER (JENNY FINKEL et al., 2005).

A Figura 1 apresenta a metodologia empregada. Para cada documento, foi extraído o conjunto das entidades nomeadas citadas no texto, relacionadas

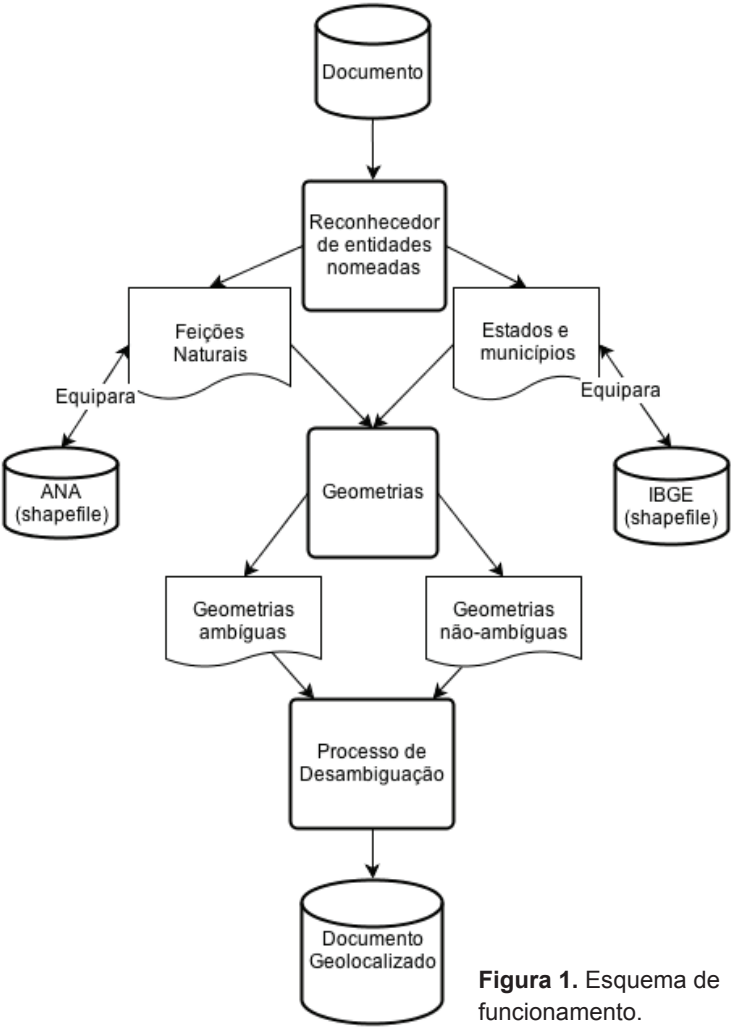

Figura 1. Esquema de funcionamento.

às feições naturais, estados e municípios. As possíveis extensões geográficas associadas a cada entidade são extraídas de arquivos de mapas no formato shapefile, isto é, o mapa da hidrografia de todos os rios brasileiros (AGÊNCIA NACIONAL DE ÁGUAS, 2005) e o mapa da malha municipal brasileira (IBGE, 2007). Dizemos que existe ambiguidade quando existe mais de uma extensão geográfica associada a um mesmo topônimo, por exemplo, no Brasil existem 4 cidades com o nome “Bonito”. O processo de desambiguação consiste em avaliar a somatória das distâncias entre as extensões.

Sejam $E = \{e_i\}$, $i = [1,n]$ o conjunto das entidades nomeadas de um documento D e $dist(g_1,g_2) = x, x \in \Re$ a menor distância entre os pontos, tomados dois a dois, de duas geometrias.

$\forall\ \boldsymbol{e_i} \exists\ G(e_i) = \{g_{ji}\}$, $G(e_i) \neq \emptyset$, $j_i = [1,m_i]$, onde $G(e_i)$ é denominado Conjunto de Geometrias Ambíguas de e_i.

Definimos G como Conjunto de Geometrias Desambiguado, onde:

$$\bar{G} = \{(\bar{g}_1, ..., \bar{g}_n)\} \mid \bar{g}_1 \in G(e_1), ..., \bar{g}_n \in G(e_n)$$

$$(\bar{g}_1, ..., \bar{g}_n) \in \bar{G} \Leftrightarrow \sum_{i \neq k} dist(\bar{g}_i, \bar{g}_k) = min\left(\sum_{i \neq k} \mathrm{dist}(g_{ji}, g_k)\right),\ k = [1,n]$$

Futuramente pretende-se estender a metodologia de modo a criar um polígono que envolva as geometrias não-ambíguas com cada uma das geometrias ambíguas, retornando o polígono envolvente de menor área, associado às geometrias, definindo assim o Conjunto de Geometrias Desambiguado.

Referências

AGÊNCIA NACIONAL DE ÁGUAS. **Rede hidrográfica brasileira**. Disponível em: <http://www.ana.gov.br/bibliotecavirtual/redeHidrografica.asp>. Acesso em: 09 set. 2013.

CARDOSO, N. Rembrandt: reconhecimento de entidades mencionadas baseado em relações análise detalhada do texto. In: ENCONTRO DO SEGUNDO HAREM (AVALIAÇÃO DE RECONHECEDORES DE ENTIDADES NOMEADAS); INTERNATIONAL CONFERENCE ON COMPUTATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 2008. [**Anais**...] Aveiro: Springer, 2008.

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.

THOMPSON REUTERS. **OpenCalais**. Disponível em: <http://www.opencalais.com/>. Acesso em: 09 set. 2013.

VARGAS R. N. P.; MOURA, M. F.; SPERANZA, E. A.; RODRIGUEZ, E.; REZENDE, S. O. The SpatialCIM methodology for spatial document coverage disambiguation and the entity recognition process aided by linguistic techniques. In: GEOSPATIAL INFORMATION AND DOCUMENTS; PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 16., 2012, Kuala Lumpur. **Workshop**... [S.l.: s.n.], 2012. Disponível em : <http://www.alice.cnptia.embrapa.br/handle/doc/948462>. Acesso em: 01 set. 2013.

Literatura recomendada

NLTK 2.0 documentation. Disponível em: <http://nltk.org/>. Acesso em: 01 set. 2013.

STANFORD NATURAL LANGUAGE PROCESSING GROUP. **Stanford Named Entity Recognizer (NER)**. Disponível em: <http://nlp.stanford.edu/software/CRF-NER.shtml>. Acesso em: 1 set. 2013.