Extração da localização geográfica em redes sociais: uma abordagem utilizando a base de dados GeoNames

Liriam M. Enamoto, Adriana C. F. Alvim, Vânia M. F. Dias
Universidade Federal do Estado do Rio de Janeiro
Av. Pasteur 458, Urca, Rio de Janeiro, RJ, Brazil
liriam.enamoto@uniriotec.br, adriana@uniriotec.br,
vania@uniriotec.br

RESUMO

As redes sociais têm sido utilizadas por pessoas de diversos países como uma ferramenta de comunicação, gerando conteúdo sobre tópicos variados e permitindo o compartilhamento de informações. A análise destas informações, quando combinadas com suas respectivas localizações geográficas, permite extrair informações valiosas das redes sociais. A localização geográfica é uma informação relevante que pode ser utilizada para monitorar doenças e epidemias, analisar preferências de consumo de produtos, analisar tendências políticas e econômicas, entre outras aplicações. No Twitter, a informação da localidade pode ser inserida manualmente no perfil do usuário ou por meio da habilitação do dispositivo GPS, o qual fornece as coordenadas geográficas latitude e longitude. Entretanto, a habilitação do GPS não tem sido adotada pela maioria dos usuários. Por outro lado, cerca de 75% dos usuários do Twitter possuem alguma informação de localidade digitada manualmente, as quais necessitam ser tratadas. O objetivo deste trabalho é propor um algoritmo para identificar o país de origem de usuários do Twitter a partir da latitude e longitude, caso o GPS esteja habilitado, ou por meio do campo *Location* informado manualmente no perfil do usuário. Neste último caso, o algoritmo possibilita extrair o país quando o usuário informar uma única localidade ou múltiplas localidades e utiliza a base de dados GeoNames como fonte de referência. Os tipos de localidades tratados são: nome de país, sigla de país, nome de cidade, nome de região administrativa e nome alternativo de localidade. Este algoritmo baseou-se no fluxo do algoritmo utilizado no trabalho de Valkanas e Gunopulos [2012], em que se dividiu a lógica nas fases de limpeza, separação de palavras ou *tokens* e na busca da localidade do usuário do Twitter com o uso das informações do GeoNames. Observou-se algumas limitações neste trabalho como, por exemplo, o não tratamento de nomes de localidades iguais que se referem a locais fisicamente diferentes, e o descarte de localidades escritas em caracteres orientais (japonês e chinês). Estas duas limitações foram abordadas na presente pesquisa. O algoritmo proposto foi desenvolvido na linguagem plpgSQL, linguagem nativa do banco de dados PostgreSQL. Foram processados dados de 349.159 usuários do Twitter coletados de novembro de 2014 a abril de 2015. Destes 349.159 usuários, 217.469 (62,28%) informaram o campo *Location* e 2.326 (1,06%) habilitaram o GPS, totalizando 219.795 usuários com alguma informação de localidade. Do total de usuários que informaram o campo *Location*, 133.471 usuários (60,73%) tiveram o país de origem identificado por meio deste algoritmo e não foi possível determinar o país de 86.324 usuários (39,27%). O algoritmo proposto permite efetuar uma posterior análise espacial, a partir de dados do Twitter, podendo ser aplicado em diversas áreas de pesquisa.

PALAVRAS CHAVE. Localização geográfica, redes sociais, Twitter.

Tópicos (OA - Outras aplicações em PO)