Exploring Multiple Evidences to Infer Users Location in Twitter

Erica Rodrigues*
Departamento de Estatística, Universidade Federal de Ouro Preto, Ouro Preto, Brazil - ericarodrigues@iceb.ufop.br

Renato Assunção
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil - assuncao@dcc.ufmg.br
Gisele L. Pappa
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil - glpappa@dcc.ufmg.br
Diogo Renno
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil - renno.diogo@gmail.com
Wagner Meira Jr.
Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil - meira@dcc.ufmg.br

Online social networks are valuable sources of information to monitor real-time events, such as earthquakes and epidemics. For this type of surveillance, users location is an essential piece of information, but a substantial number of users choose not to disclose their geographical information. However, characteristics of the users behavior, such as the friends they associate with and the types of messages published may hint on their spatial location. In this paper, we present a method to infer the spatial location of Twitter users. Unlike the approaches proposed so far, we incorporate two sources of information to learn geographical position: the text posted by users and their friendship network. We propose a probabilistic approach that jointly models the geographical labels and Twitter texts of users organized in the form of a graph representing the friendship network. We use the Markov random field probability model to represent the network and learning is carried out through a Markov chain Monte Carlo simulation technique to approximate the posterior probability distribution of the missing geographical labels. We show the accuracy of the model in a large dataset of Twitter users, where the ground truth is the location given by the GPS position. The method is evaluated and compared to two baseline algorithms that employ either of these two types of information. The results obtained are significantly better than those of the baseline methods.

Keywords: Network Learning; Geographic Targeting; Geolocation Estimation.