Twitter information retrieval using web crawlers

Luiz D. R. França[1]

[1]Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brazil

luizdaniel.r.f@gmail.com

Abstract. *Twitter is a microblogging service that allows the users share ideas with short messages. The twites (the messages sent on Twitter are called twittes) are limited in length by 140 characters which makes the information extraction and retrieval difficult. This project will extend the RetriBlog framework [Ferreira et al. 2013] to make this process easier.*

1. Introduction

Twitter is a social network that allows people to exchange 140-character messages. Nowadays many people use twitter to share their ideas and opinions, even companies use it to show their products and interact with their costumers. With the growth of social networks like Twitter, a huge amount of data is produced everyday, in a single second five hundreds million twitters are sent [Stats]. This project is meant to develop a solution for retrieve information from Twitter using web crawlers.

2. Problem

As stated earlier, twittes are 140-character messages exchanged between hundreds of million users around the world. The small length of the twittes makes it hard to extract relevant information using traditional methods [Sriram et al. 2010]. With so much information uploaded everyday, it's very hard to get the most relevant twittes and extract information from it. The fact that the twittes are only 140-characters long make it even harder. With messages full of abbreviations and slangs.

3. Reason

The information retrieved from Twitter has many applications like predict stock market [Bollen et al. 2011] and monitor political sentiment [Bermingham and Smeaton 2011]. This project is relevant because it will make easier to explore the richness of information on Twitter. Futhermore it's going to allow the retrieval of relevant information that otherwise would be very difficult to find due to the vastness of the contents uploaded everyday on Twitter.

4. Goal

This project proposal is meant to develop a framework that can retrieve information from Twitter. We are going to extend the RetriBlog framework [Ferreira et al. 2013]. that is a framework to extract.

5. Related Jobs

These are some works related to information retrieval and extraction of information:

- Short Text Classification in Twitter to Improve Information Filtering [Sriram et al. 2010]
- Managing the Acronym/Expansion Identification Process for Text-Mining Applications [Roche and Prince 2008]
- Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments [Gimpel et al. 2011]

References

Bermingham, A. and Smeaton, A. F. (2011). On using twitter to monitor political sentiment and predict election results. *Workshop at the International Joint Conference for Natural Language Processing*.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*.

Ferreira, R., Freitas, F., Brito, P., Melo, J., Lima, R., and Costa, E. (2013). Retriblog: An architecture-centered framework for developing blog crawlers. *Expert Systems with Application*.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Roche, M. and Prince, V. (2008). Managing the acronym/expansion identification process for text-mining applications. *Int J Software Informatics*.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. *ACM*.

Stats, I. L. Twitter usage statistics.