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 day five hundreds million twitters are sent [Stats]. With so much information uploaded everyday, it's very hard to get the most relevant twittes and extract information from it. Messages full of abbreviations and slang and small length of the messages on Twitter makes it even harder to extract relevant information using traditional methods [Sriram et al. 2010]. 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. 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].

2. Project Design

- Twitter Crawler: the crawler will run through the twitter pages and store the links to later be saved. - Store Documents: this part will get the links from the crawler and it will save the twittes. - Preprocessing: - Acronyms Expansion: as twittes are only 140-characters long, many people use acronyms on their texts. This module will expand those acronyms to its normal form. - HTML Clean: this part is responsible for removing the html tags - English Stemming: this part will reduce the words to it's lemma. - Text Expansion: this module will expand the text. - Indexing: this module will create an index of the twittes.

3. Related Jobs

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

- RetriBlog: An architecture-centered framework for developing blog crawlers [Ferreira et al. 2013]
- 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]
- Traffic Condition Information Extraction and Visualization from Social Media Twitter for Android Mobile Application [Endarnoto et al. 2011]
- Open domain event extraction from twitter [Ritter et al. 2012]

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*.

Endarnoto, S. K., Pradipta, S., Nugroho, A. S., and Purnama, J. (2011). Traffic condition information extraction visualization from social media twitter for android mobile application. *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*.

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*.

Ritter, A., Mausam, Etzioni, O., and Clark, S. (2012). Open domain event extraction from twitter. *KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*.

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.