DOI: 10.20396/revpibic2720191882

XXVII Congresso de Iniciação Científica
Unicamp
16 a 18 de outubro de 2019 - Campinas | Brasil

Monitoring of events in social networks for data collection, information analysis and temporal synchronization

José Nascimento*, Anderson Rocha

Abstract

To collect social media footprints about an event, the usual gathering technique consists on a series of keywords, and this approach invariably is not able to collect all items that are actually connected to the event. In order to minimize this drawback, we propose herein the design and development of a data collection technique leveraged by query expansion for social media items. The objective is to augment the set of social media items related to the event.

Key words:
Social network, query expansion, event tracking

Introduction

In the sense of monitoring events, Schinas et al. developed the **Social Tracker** system [1], which allows us to collect items from different social networks (e.g. Twitter, Youtube), by a *keyword-based search.* In this system, a set of keywords and collected items (which text matches these keywords) is called a *collection.*

From this perspective, this work focus in researching *query expansion techniques* to increase the set of items of a given collection. **The aim of a query expansion is to discover new keywords related to the event that were not previously defined, and redo the initial query added with the discovered keywords.**

Results and Discussion

Given the collapse of a bridge over the Moju River in Brazil's Pará state [2], that occured on April 6, 2019, we may define for this event the *collection_1* using the keywords "*ponte pará"* and "*ponte paraense".* After a while, 880 items have been collected. Then we may apply query expansion techniques.

The system already stores tags of each item. In this sense, we developed the ***Tags Method***, a naive procedure which ranks all tags based on their frequencies, then suggests the most common ones.

Three improvements in this method were made. Firstly, we implement a feature that here are going to be called ***variance analysis***. It consists in grouping tags which only differ diacritical marks, then count them as one item only (summing the appearances), and finally suggesting each word as a possible new keyword.

Then, we start the development of a ***query expansion stopword list***. A suggested tag might be too much general, and collect considerably more non-related items than related items. The list is intended to work around this problem. It contains the most common tags in portuguese-language social media items that we collected (summing all collected collections we have).

Finally, names of places are very common in event-related items. However, they are also too general, and may bring many non-related items. In this sense we generate a ***geographical name stopword list***, based on the *GeoNames geographical database* [3]. This list contains the name of each country, state or region related to a given nation (In this case, we are going to use Brazil's list).

Applying the Tags Method in collection_1 with these three improvements suggests the keywords in chart 1.

Chart 1. Collection_1's twenty most common tags. Italic indicates variance analysis; in red, tags in the query expansion stopword list; in green, tags in the geographical name stopword list. In purple, tags in both lists. Finally, in blue are the ones added to the collection.

para	***pará***	**rio moju**	**ponte**
noticias	***notícias***	**balsa**	**brasil**
acidente	**ponte caiu**	***alca viaria***	***alça viaria***
alça viária	***ponte cai no para***	***ponte cai no pará***	**r7**
moju	**ponte do moju**	**ponte rio moju**	**jornal**

After adding the new keyword to collection_1, it has 2718 items. The augmentation is expected to be proportionally higher in bigger events.

Conclusions

The improvements were developed according to problems we noticed when executing the *Tags Method*. The keywords that were added is expected to hardly ever collect non related items (possible exceptions: *ponte* and *balsa).* On the other hand, a method to strictly evaluate the quality of a query expansion is still to be developed.

Furthermore, words like "*pará"* and "*moju"* are related to the event, and should not be just ruled out. In this sense, one of the next steps of this work is researching the addition of these type of keyword in pairs.

Acknowledgement

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

[1] https://github.com/MKLab-ITI/mmdemo-dockerized
[2] https://www.reuters.com/article/brazil-grains-bridgeaccident/brazil-bridge-collapse-could-affect-grain-shipments-in-north-idUSL1N21O05Z - Retrieved in 2019-04-04

[3] GeoNames. http://geonames.org/. Retrieved in 2019-07-09