SocialSOM: Topic Detection on Twitter by Organizing Tweets on User Similarity

Bernardo Simões 20-25 páginas aproximadamente

Technical University of Lisbon - Taguspark Campus,
Av. Prof. Doutor Aníbal Cavaco Silva — 2744-016 Porto Salvo, Portugal
bernardo.simoes@ist.ul.pt
http://www.ist.utl.pt/en/

Abstract. *70 and at most 150 words, What did I do, in a nutshell?, summarize the paper, should be written last , very short context ,what the objectives of the study were*

Keywords: topic detection, twitter, self-organizing maps, classification, clustering

1 Introduction

With the evolution of social networks websites like Facebook or Twitter throughout the last couple of years, the amount of pertinent content about a specif issue is increasing dramatically, which calls for new ways to make sense and catalog this data. In the other hand find topic sensitive information on social networks is extremely complicated due to the fact that documents have very little content, slang vocabulary and orthographically mistakes or abbreviations.

The value of data presented in sites like Facebook or Twitter as proven its value in papers like "Predicting the future with social media" where it is possible to predict with high precision the value of a movie box office weeks before it debuts.

This project will focus on Topic Detection on Twitter by using a new approach that will not only take in consideration the words in the corpus, but will also take in account the social network to which a tweet belongs to in order to categorize it using the concept of homophily that has been proven in past that is applicable to social networks.

Will be described the objectives of this project, at 2 we will talk about the state of the art solutions related not only to topic detection but also to twitter data analysis and Self-Organizing Maps.In section 3 Architecture of the purposed solution and finally at section 4 it will be discussed how to evaluate results achieved.

1.1 Objectives

The objective of this project is clear, finding topics on Tweets by analyzing their corpus specific characteristics, like number of characters in a tweet, hashtag, "was retweeted", etc.. And contextualize the social network evolving the person that did the tweet.

After characterizing the tweet with information just described, we will use the unsupervised learning clustering technique Self-organizing maps in order to organize the tweets in clusters of topics. Afterwards it will be needed to categorize the clusters in order to know which topic they belong to.

Lastly the resulting topic clusters will be publicly accessible through a website to everybody that visits it.

2 Related Work

- What did we know about the problem before I did this study?
- What did we do different from previous works?
- Discuss the relevant primary research literature
- Works should be organized by their relevant characteristics
- Comment on why it is relevant for your work
- Comment on what your work does differentely

2.1 Clustering and Self Organizing Maps

2.2 Topic Detection on Twitter

3 Architecture

In this project we are going to apply Self Organizing Maps in order to detect clusters of Topics on Twitter.

3.1 Data Gathering

In order to retrieve data from Twitter, we will be using a ruby library called Twitter Stream [1] , that enables the user to download and inspect the twitter stream of tweets. As the data is gathered it will be stored in a MongoDB database for posterior analyses. As the twitter stream is stored, another function will interact with the twitter API in order to retrieve information from a user profile and relate him with other users by analyzing his followers and who the user is following. In the end of the data-gathering process it will be possible to query the database for:

- Tweets from a user.
- Query tweets for hashtag.
- Query users followers and who he is following.
- Query for tweets that shared the same URI

Tweets will be cathegorized in:

- News Accounts
  - Accounts with a lot of followers
- Profile custumization
- Average number of tweets a day with uri (might suggest spam)

- How am I gonna solve the problem?
- Describe the work that will be done

4 Evaluation Metrics

- How am I gonna evaluate my work?

4.1 Evaluation Criteria by Teachers

- Ability to understand the research problem
- Clear and well defined goals
- Description of the different approaches explored
- Ability to relate the state-of-the-art with the research theme Work methodology and adequate planning for the next stage Organization and quality of the written document
- Inclusion and completeness of updated and appropriate references Oral presentation and discussion

References

1. Steve Agalloco. TweetStream, simple Ruby access to Twitter's Streaming API. 2013.