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

Why use topic detection Why use clustering Why use SOMs

Setting the context: Twitter is the most popular microblogging platform that enables users to share their own thoughts in less than 140 characters. Through out the years twitter evolved from a simple platform to share what a person is doing into an message broadcaster. Nowadays twitter is used in every kind of areas, from journalism to political campaigns. The way users engage in the social network was enhanced by adding Hashtags a way to tag a certain topic. Sharing web content through urls in order to direct link a certain article or blog post.Responses can be made to a user twit in order to engage users to talk directly about a certain issue and retweets a way to broadcast a tweet from another user.

clearly communicate what you want to discover: In this report we will introduce the concept of using Self Organizing Maps in order to cluster tweets based on the Social Network. What we hope to acheive is a new way to find topic detection based on the concept of Homophily which states that people tend to associate and bond with similar others. If people tend to follow other people with the same interests it is possible that groups of people tend to tweet about certain topics.

In the end of the project, it will be possible to use a web interface to see topic-clusters based on the Twitter social interactions.

1.1 Self Organizing Maps Usage

☐ Set the context;

- ☐ explain the situation;
- ☐ State why the main idea is important;
- ☐ provide general information about the main idea
- ☐ clearly communicate what you want to discover why you are interested in the topic
- ☐ Outline the structure

2 Objectives

3 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 differently

3.1 Clustering and Self Organizing Maps

Cheong at al[2] analyzed user profile characteristics based on a certain topic using SOM. On three different topics, "IPhone Software Launch", "Obamas Foreign Policy" an "Iran Election" they could find distinct types of users twitting:

- "IPhone Software Launch"
  - Major twits where from male users, with accounts greater than 90 days, coming from countries where iPhone was marketed, with high adoption of blogs or social media sites.
  - Twits with higher ratio of followers to followees with high frequency of twitter posts per day, twits with links to shared content, no country or gender specificity. Typically a news aggregator or news organization.
  - One day twit account, with unpopular social connections, lacking profile customization, frequently posts more than 50 tweets daily with URIs. Characterizing a typical spammer.

- "Iran Election"
  - Recently registered, Iranian web-based twitter users, frequent patterns of replying
  - Users from every where in the world, long message sizes. Users trying to raise awareness
  - Users with accounts older than 3 months, contribute sparingly to Twitter, but have a high usage of other social media sites.
  - Variance in Twitter account and nationality, who frequently posted URL links in messages

- Obama's foreign policy

- American residents discussing the topic, accounts more than three months old, their messages are almost always long, and their messaging style is focused towards replies. Describing users talking about the issue
- Users with many followers, predominantly US males, URI links in their messages. Which describes news sources and opinion leaders.
- Mainly new accounts from everywhere around the world which arises suspicions of marketing/opinion spam.

It is possible to see based on their results, that there is direct relation between a twitter profile characteristics and the content produced by the user. Even though in the paper is not described the amount of signals used, it is possible to determine the following:

– The time when an account was created.
– If a tweets is used to start a thread.
– The size of a tweet.
– Localization of the tweet.
– Number of followers and followees.
– Number of contributions to twitter.
– Profile customization.
– Popularity of the connections (number of followers of the followers)
– Content of the URI´s
– Number of tweets a day.

3.2 Topic Detection on Twitter

4 Architecture

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

4.1 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

4.2 Data Characterization

Depois de se tirar twits a partir da API de streaming do mesmo, vamos contar as palavras mais utilizadas na rede social depois de se remover stop words, palavrões e abreviaturas socias (como lol, omg, brb e combinações das mesmas). As palavras com maior ocurrencia vão ser utilizadas como identificadores num tweet. De seguida para cada utilizador conta-se a quantidade de vezes que cada um mencionou cada uma das palavras com maior frequencia no twitter, desta forma identificamos o tipo de conteudo que um utilizador tem tendecia a produzir. De forma a criarmos a representação de rede social iremos passar 30 por cento das palavras mais escritas de um utilizador para os seus followers e acrescentalas ao raking do mesmo.

Os SOM vão organizar os teweets com base na representação social de um utilizador, deste modo esperamos encontrar núcleos de tweets com utilizadores com formas de escrita e interesses parecidos que faça com que os seus tweets sejam sobre tópico similares.

De forma a se fazer topic detection iremos pegar em palavras chave de um determinado tópico e porcurar em clusters que apresentem o maior numero de utlização dessas palavras.

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

5 Evaluation Metrics

- How am I gonna evaluate my work?

5.1 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.
2. Marc Cheong and Vincent Lee. A Study on Detecting Patterns in Twitter Intra-topic User and Message Clustering. *2010 20th International Conference on Pattern Recognition*, pages 3125–3128, August 2010.