Visualizing the contributions of a user in the Social Web

AUTHORS:
Maria Schmidt
Maria-Stamatoula Karavolia

UNIVERSITY:
University of Fribourg
COURSE NAME:
Web Monitoring and Analysis
SUPERVISOR:
Aleksandar Drobnjak – Research Assistant

May 21, 2016

Contents

List of Figures

1 Introduction

In the recent decades, the notion of social networks have attracted the interest of researchers and the curiosity of the social behavioral sciences, since there is a wide spread usage of the internet by people all over the world that interact with each other and exchange web content through numerous online communities such as Facebook[1], Twitter[2] and many other. An interesting aspect in social networks is the visualization of the contributions of a user posts and their impact across different platforms. Over time the contributions of an user in social networks is growing and this creates the need to have a simple overview of the data by visualizing it. Data visualization offers a quick way to present the data in a way that can reveal valuable hidden insights. Thus, through the visualization, users can easily understand what are the hot posts, that is those posts are able to attract a greater attention or interest.

[1] http://www.facebook.com
[2] http://www.twitter.com

2 Motivation

klgdfkgdkfgdkljgkldf

3 Overview

klgdfkgdkfgdkljgkldf

3.1 Research questions

1. How can we visualize the contribution of a user in social networks?
2. How can we interrelate the posts published by the same user?
3. How can we find similar posts in different social networks?

3.2 Goals of the project

kdsasdjkakjdkajasjkl

4 Prototype

In this section we provide a detailed description of our framework, the data collection and the technologies that used in the front and back ends.

4.1 Framework

Our web application is based on Django[3], which is a free and open-source web framework, written in Python, and follows the model–view–controller (MVC) architectural pattern.

4.2 Data Collection

Our data collection consists of real-world data from Facebook and Twitter and focus on 28 public persons such as, politicians and athletes because they tend to post the same content on Twitter and Facebook more than a normal user.

4.2.1 Retrieve data from Facebook

4.2.2 Retrieve data from Twitter

The data was obtained by quering the timeline API of Twitter with the username of each person related to politicians and athletes. For this procedure we used the Tweepy[4], which is a Python library for accessing the Twitter API. We were able to collect a fixed number of tweets because Twitter only allows access to a users most recent 3240 tweets. The attributes of the data along with their definitions are displayed in Table 1.

Table 1: Description of the attributes of Twitter data

Attribute	Description
ID	The id of the twitter post
date	The date when the tweet was posted
text	The text of the tweet
likes	The number of likes of a tweet

[3] `https://www.djangoproject.com/`
[4] `http://www.tweepy.org/`

4.3 Front-end

4.4 Back-end

5 Aspects of user visualization

5.1 Amount of contribution

5.1.1 Theory

5.1.2 Visualization

5.2 Topic Models

5.2.1 Latent dirichlet allocation

Latent Dirichlet allocation (LDA) [1, 2, 6] is a unsupervised machine learning technique that discovers latent topics from a corpus of documents so that the documents can then be assigned automatically into appropriate topics. New documents can also be classified into topics based on these latent topics. More specifically, in LDA, each document is represented as a mixture of various topics, where each topic is a mixture of words. These mixtures are represented by P(topic|document) for all topics and documents and P(word|topic) for all words in the vocabulary. Each word may occur in several different topics with a different probability, and each document is assumed to be characterized by a particular set of topics.

5.2.2 Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) is an unsupervised family of algorithms that simultaneously perform dimension reduction and clustering.NMF was first introduced by Paatero and Tapper [5] as positive matrix factorization and subsequently popularized by Lee and Seung [4].

5.2.3 Compare LDA and NFM

5.2.4 Visualization

5.3 Impact of post

5.3.1 Theory

5.3.2 Cosine Similarity

For creating the user timeline, we need to find similar posts that a user share both in Twitter and Facebook. For this purpose, we calculate the cosine similarity metric.The cosine similarity [3] between two vectors (or two documents on the Vector Space) is a

measure that calculates the cosine of the angle between them. This metric is a measurement of orientation and not magnitude, it can be seen as a comparison between documents on a normalized space because we're not taking into the consideration only the magnitude of each word count (tf-idf) of each document, but the angle between the documents.

Given two posts t_a and t_b, their cosine similarity is

$$\cos(\mathbf{t_a}, \mathbf{t_b}) = \frac{\mathbf{t_a t_b}}{\|\mathbf{t_a}\|\|\mathbf{b}\|} \tag{1}$$

where t_a and t_b are m-dimensional vectors over the term set $T = t_1, ..., t_m$. Each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between $[0, 1]$.

We predefine a threshold to accept two similar posts to have similarity at least 60%. This portion lets us a great amount of similar posts and also recognize twitter posts that have urls, hashtags and mentions.

5.3.3 Visualization

5.3.4 Other possibilities to visualize text corpus

6 Conclusion and future work

jkfsjkfjksfkdjkfdk

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] G. Heinrich. Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005.

[3] A. Huang. Similarity measures for text document clustering. pages 49–56, 2008.

[4] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[5] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[6] M. Steyvers and T. Griffiths. *Probabilistic topic models*, volume 427. Handbook of Latent Semantic Analysis, 2007.