ABSTRACT

SP14:332:438 SECTION 1 Capstone Design – Project Report

HEALTH MONITORING ANALYTICS

Understanding the Impacts of Individual Responsibility of Communities through Twitter

Xiaoyu Yu  keyxy@gmail.com
Zhan Chen  johnalwaysyoung@gmail.com
Yue Song  ys394@scarletmail.rutgers.edu

https://sites.google.com/site/heathmonitoringanalytics/
Contents

Abstract
Introduction
Approach
Result
Sustainability Analysis
Conclusion
Bibliography
Individual Contribution

All members contribute equally

- Twitter API, Data Mining, Main documentation writing.
  - Xiaoyu Yu
- Google Maps, build heat map, poster design
  - Yue Song
- Webpage design, Data analysis, Data base and server setup,
  - Zhan Chen
Abstract

In this project, we develop an application which can collect data from twitter and search the key words related with exercise and furthermore visualize the data on google maps. The application is customized with personal account, which can track customs’ exercising activity, determine the health situation and give advices to help develop a healthy lifestyle. In this way, people will be more motivated to do more exercise thus achieve the goal of staying healthy.

The tweets and search results are collected using streaming API, including the location identified through geotagging information and their profile. We use Mongo DB as one example of a NoSQL implementation to deal with the large volume of Tweets, Tweeters, and network information. By using Mongo DB, we can store data from Twitter’s APIs and to search it for specific information. We select topics in the text and build a hashtag automatically. To analyze of emotional factor of the data, we do a two-layer-filter key searching, and re-arrange the data. Finally we visualized the data with network, temporal, geo-spatial and textual information.
Software architecture

Software Interface

Heat Map

Statistic Chart

Data Analysis

Database

Storage

Twitter Server

visualization
Introduction

Twitter is an online social networking and people can conveniently share their opinions about things there. The words people upload onto twitter are called tweets. The number of words of every tweet is limited in 140, which make people express their opinion and thought in a more precise way, in terms of helping the reader in the twitter more easily catch the point of others’ words. Nowadays, people posted more than 340 million tweets per day. It has created a large platform that people from all over the world share and discussion their opinion and thought.

But why can't tweet be a tool for us to improve people’s health condition? Since human beings are social creatures, the actions of the surrounding people definitely influences the actions of the individual. Think about that most people around you are aficionado of sport activity, it is hard for you to prevent from that influence. Promoting people to exercise is good in most situations for reasons as follows:

1. Most people don't have enough exercise rather than excess exercise.
2. Human beings are not like machines, sometimes our body cannot react to us immediately thus it is often too late when some decease comes to us.
3. It is much easier to take actions to prevent the situations of body’s disorder rather than trying to figure it out when you are caught by the deceases.

Think about we analyze the data collected from twitter and show people how the people surrounding them love doing sports, it is not hard to imagine how they will be changed by the way of other people’s lifestyle. It is of high possibility that the people will be prompted to do more exercise when they see the analysis.

In this project, we analyze the conditions of people’s health mainly based on the contents of the tweets people posted in twitter. We do this based on the following reason:

1. As the content described above, there are enough amount number of tweets posted every day. What’s more, people can share their current location while posting their tweets, which makes us able to combine both the locations and the contents and do the reasonable analysis.
2. We do the analysis mainly based on the topics on people's tweets. For example, while people posted a tweet which contains words related to sports, say, gym, basketball, we count this tweet as one positive factor of health.

3. While we analyze the tweets, it worth considering about the similarity of the topics because it is very important in the process of our analysis. Thus we try to introduce the LDA (Latent Dirichlet allocation). It is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. While using LDA, we can do the analysis in a more precise way thus get more accurate analysis result.

This project is mainly based on analyzing the topics of tweets posted on the internet and predicting the health condition of people in different areas. In this way, we can both motive people to exercise more to keep health and predict the sports marketability of different area.
CRAWLING TWITTER DATA (USING THE TWITTER API)

Twitter uses Open Authentication (OAuth) to keep users’ information safe from third parties. Information we could retrieve from Twitter includes users’ profile, network (followers, friends), and most importantly their tweets with relevant hashtag metadata.
Profile data we need includes users handle, location, URL (which leads to a more specific profile), social networking activities between other users, and the amount of tweets posted as well as their registration date. All of these can be achieved using the GetProfile method which uses a unique identifier to retrieve one’s profile.

Data of one’s followers and friends can be crawled using followers/list and friends/list respectively. A maximum of 15 followers can be searched out as a list while using Twitter handles. For friends, this number can be as many as 20 in a list. Moreover, the exact location of users can be identified through geo tagging information and their profile.

Figure 2 Tweet data case

STORING TWITTER DATA:

Managing the huge volumes of aggregated data is quite a big problem (typical with big data). We will use MongoDB as one example of a NoSQL implementation to deal with the large volume of Tweets, Tweeters, and network information.

MongoDB is an open-source database used by companies of all sizes, across all industries and for a wide variety of applications. It is an agile database that allows schemas to change quickly as applications evolve, while still providing the functionality developers expect from traditional databases, such as secondary indexes, a full query language and strict consistency. MongoDB is built for scalability, performance and high availability, scaling from single server deployments to large, complex multi-site architectures. By leveraging in-memory computing, MongoDB provides high
performance for both reads and writes. MongoDB’s native replication and automated failover enable enterprise-grade reliability and operational flexibility.

By using MongoDB, we can store data in JSON-style, which could make it easy to store data from Twitter’s APIs and to search it for specific information. JSON (JavaScript Object Notation) is a light weighted data-interchange format which is easy for programmer to read and write. Besides, JSON format is suitable for machine to parse and generate thus is widely used in transmit data between a server and web application. For MongoDB, its fast query speed is also appreciated. To implement data storage, we will add tweets to the collection first. We will optimize the collections for queries and create the indexes. Then, retrieve all documents in a collection. After that, filter documents and find the most relevant tweets. Last but not least, identify the most popular users, which will be important for analysis.

Each tweet contains several JSON objects, some of which important information are as below:

<table>
<thead>
<tr>
<th>key</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>Contents of the tweet</td>
</tr>
<tr>
<td>id</td>
<td>User’s ID</td>
</tr>
<tr>
<td>created_time</td>
<td>Creating time of the tweet</td>
</tr>
<tr>
<td>location</td>
<td>The location user filled in his/her profile</td>
</tr>
<tr>
<td>user</td>
<td>User’s information (name, sex and etc.)</td>
</tr>
<tr>
<td>position</td>
<td>Longitude and latitude of the user when user posted tweet</td>
</tr>
</tbody>
</table>

**Figure 3**

**ANALYZING TWITTER DATA:**

The basic idea of analyzing data is to implement a type of machine learning. When analyzing Twitter data, we will use proper network measures to find the important topics and information that are relative to our project. First, we will explore a model
that exploits the links between the entities to find key players in the data and understand the text to find relative information. Second, we will explore topic model that find natural topics in the text. By using LDA algorithm, we will discover select topics in the text and build a hashtag automatically. How to analyze of emotional factor of the data? We use multiple processes to achieve this goal.

i. First filter
   We search the key related to our topic, which is sports (e.g. Gym, sport, exercise, etc.) and get the “topic” chain which include the key;

ii. Second filter
    We search of key with emotions (e.g. Like, hate, display, watch, etc.) with the "topic" taken out, and give every emotional key a weigh and then compute the total weigh of keys, which represent the overall healthy condition of that field.)

iii. Re-arrange
    The next step is to re-arrange the data we get after the emotional analysis, this part is mainly divided into two parts, delete the useless info for each member data, and re-sort the data based on the list we get from emotional analysis.

iv. Format changing
    A general format which would be accepted by the following process is one of the core in the data analysis part, as far as our project, JSON is the mainly kind of document that widely used in mongo, however, it is not a general format for such a lot of software, how to transfer it with its whole data is the key part for current process. As several tries, we temperately choose csv as the ideal format.

v. Double check and re-sort new document
Data losing or sequence changes are the mainly risks when we transfer into another format with such a great deal of data, especially in one situation that it change the array member, which means the new arrays do not match anymore. Therefore, double check is necessary.

![Figure 4 Regional real-time statistic](image)

*Figure 4 Regional real-time statistic*

When choosing a certain area (eg. New York) and a key word (football), we can derive the amounts of tweets.
We extract a time period in April 11th using the same area and hashtag to show the time stamp of tweets.

Figure 6 Time stamp of tweets

VISUALIZING TWITTER DATA

The program is easily to be divided into two general parts, mining and visualize, therefore, realize a visual and vivid result is very important in our product. After analyzing the data, we got the information and ratio that we need. Based on those data, we make the display, mainly into the charts, and the maps. The charts parts is much
easier, we only need to call those ratio and name of the region. For the maps (Google Map and Google Earth), we have to get a user protocol to get the permission to call those functions in our product, after that, we transfer our data into those API.

This function can help users find out the location of other users on Google Earth. Our system is a real time system, so, if somebody just posted a tweet which related to the key words on twitter, there will be a yellow spot appears on map at once and we can also in store and extract historical data from out data base to show the overall situation. The function's purpose is to virtualize the result of analytical data, the users can find out the result on map directly.

According to the historical data, we can deal with them and after that, we can show a heat map below, the different color of heat map represents the different density of tweets. In other words, in this specific area, if the density of related tweets is high, then the color of this area is darker. And if in that area, few people posted related tweets, the color of the area is light.

This function 's purpose is to virtualize the density of different areas, the users can easily find out which area has more people participate in the exercise and post related tweets on twitter.
EXTRA FUNCTIONS

After molding the product, the next point is to make it better, what we can do to improve it? Right now, we added the user login function to save user information. (Such as display it on the Google map in personal webpage.) And we also improve some suggestions to different situations. Besides that, considered that we would improve and increase our data as time goes by, we call Google search functions on our webpage, it could realize the local search, and it will connect with Twitter official, in the future, we just input a keyword, we can get the health status all over the world.

Figure 4 Heat map
Result

When login the webpage, users could indicate the location his prefer to get the information, or display the area set as default. Via zooming in or out, and drag the map, users can obtain data in any interest area.

![Software interface](image)

**Figure 5 Software interface**

By using this application, users can get real-time information based on the tweets collected and display on Googlemap, showing that where and when other people talk about sports, which partly regarded as they doing sports, and the relevant topic such as healthy diet, daily routine and mental health.

Meanwhile, registers is encouraged to build personal wellness dairy which records the exercise situation and physical condition, to support the further analyst and get the feedback.

Finally, the feedback page is available to see, a brief report base on the data provided by user and crawledin tweets.
Figure 6 Result page
Sustainability Analysis

Since our project is totally designed by software, only the energy such as electricity and expendable supplies such as paper are consumed when it is designed and produced. When it is devoted into the market, it only needs server to handle the data extracted from the Twitter API and do the analysis. Thus our software is environmental friendly and can be easily applied to market.
Conclusion

In this project, we realize the following functions:

1. Realize the heat map on Google Map and Google Earth, showing the frequency of the topics related with exercising activities in the certain area. From the heat map, users can obtain a straightforward impression of the health conditions of the people in that certain area.

2. Provide a personalized report of health condition. Users need to sign in to see their personal, in which provide the users the locations of their exercising activities, and the specific sport they do at specific date. Additionally, the report gives suggestions to the users after analyze the personal data of the users. In this way, we give users a convenient way of observing their health condition and provide them a direction of getting healthier.
Bibliography

[1] Software Engineering Project: Health Monitoring Analytics
http://www.ece.rutgers.edu/~marsic/books/SE/projects/HealthMonitor/analytics.html

[2] The project of Health Monitoring Analytics (Gradeigh D. Clark, Xianyi Gao, Rui Xu, Et Al., 2013)

http://tweettracker.fulton.asu.edu/tda/


[10] Frank Bentley." Investigating the Place Categories Where Location-Based Services Are Used.”