Social Network Data Mining Using Natural Language Processing and Density Based Clustering

David Khanaferov, Christopher Luc and Taehyung ("George") Wang
Department of Computer Science
California State University, Northridge
Northridge California, 91330
twang@csun.edu

***Abstract*- There is a growing need to make sense of all the raw data available on the Internet; hence, the purpose of this study is to explore the capabilities of data mining algorithms applied to social networks. We propose a system to mine public Twitter data for information relevant to obesity and health as an initial case study. This paper details the findings of our project and critiques the use of social networks for data mining purposes.**

Keywords-data mining; social network; clustering, NLP; sentiment analysis

I. INTRODUCTION

The amount of data generated by social networks and public access to subsets of that data make this domain a great candidate for complex data mining. For this paper we collected unstructured data from one of the most popular social networks, Twitter. We collected and mined user generated content containing patterns related to healthcare and obesity. Our goal was to demonstrate a practical approach to solving an alarming healthcare issue through a systematic, computational approach centered on mining useful patterns out of public data.

II. BACKGROUND

One of the problems associated with mining social networking data is the type of data available for mining. The bulk of most social network data is either semi-structured or completely unstructured. Currently, Twitter's public API [1] provides access to raw user generated data in the form of messages, status updates, and comments.

A. Natural Language Processing

Natural Language Processing (NLP) is an industry term for algorithms designed to take a document consisting of symbols and deduce associated semantics [2]. NLP is an active research field and many tools have been developed by prominent organizations that allow for semantic processing. We were focused on a subset of NLP for the purposes of this study, specifically text mining. Also, for convenience, we chose to focus on the English language.

B. Sentiment Analysis

In this study, we propose a clustering algorithm to identify groups of user entries (transactions) with similar properties. Sentiment is a measure of negativity towards the search subject in our model. Using the available tools, we propose to classify transactions on three dimensions: location, query relevance, and sentiment. We provide an implementation of the clustering algorithm for healthcare informatics and more specifically the topic of obesity and health as an example of our system.

III. PROPOSED SYSTEM ARCHITECTURE

We proposed a system architecture for a data warehouse to execute online analytical mining operations for mining social network data. Serving as a data layer for clustering algorithms mining processing, the data warehouse system consists of three distinct layers. The data layer is implemented as an OLTP relational database, virtual data warehouse, and visualization which is further broken down into application cache and the visualization algorithms. The lowest layer is implemented by an open source MySQL server. The virtual data warehouse layer is implemented as a set of SQL views that dynamically aggregates carefully selected data sets into a star schema [4]. The constellation schema for the proposed data warehouse consists of 3 dimension tables: Location, Date, and Tweet Transactions. We refer to a tweet transaction as a set of keywords, mined by the NLP tools from a tweet. A tweet transaction serves as the central point of focus for all further operations.

978-1-4799-4003-5/14 $31.00 © 2014 IEEE
DOI 10.1109/ICSC.2014.48

IV. Data Collection

Data was collected from Twitter by using the public Twitter API. For this project we had prioritized a list of twenty five query keywords relevant to obesity to query the Twitter API against. Two applications were written to request and save tweet and user information. Both applications queried the Twitter API, formatted the results, and inserted them into a MySQL database.

A. Data Cleaning

One of the most important tasks in successfully mining social networks is data cleaning. We took an incremental approach to data cleaning, filtering out unwanted data at each step using various NLP tools.

The first step in the data cleaning process was removing data with missing parameters. In a subsequent step, we needed to convert user location text strings into usable latitude and longitude data. Next, we needed to create the tweet transaction records. We use the term tweet transactions to identify those that are cleaned, structured, and parsed from raw tweets. A tweet transaction consists of several pieces: sentiment value and a set of keyword IDs extracted from the tweet content. Each keyword id represents a keyword object stored in the keywords table.

Another major task in creating transactions is extracting useful (interesting/meaningful) keywords, those that can affect context, from tweets and assigning them to transactions. The final step in the cleaning process for keywords was identifying semantic and morphological similarity and normalizing similar keywords to a consistent state.

B. Data Standardization

To ensure that any data mining algorithms chosen to analyze our data set performed accurately, we followed the data cleaning phase of the project with a data standardization phase. In this phase the cleaned data had to be normalized through simple mathematical functions.

V. Implementation

Due to the random nature of the raw data mined and the exclusion of a training set of data, it was essential to use clustering since the learning process was unsupervised. Clustering being the only possible way to find out patterns because it arrange similar sets of elements near one another for grouping. Clusters were plotted on a 4 dimensional space where a vector of four elements was used to determine the location of any point in that space. As a consequence of the uncertainty present in the data collected, a density based clustering algorithm was selected. DBSCAN is a type of density based clustering algorithm and was chosen to implement the clustering for this study. The output of the algorithm was a set of clusters which was then used for visualization.

VI. Results

The output of the DBSCAN algorithm is a set of clusters where every cluster consist of a sets of transactions, and each transaction includes a set of search associated terms. To visualize the clusters created by the DBSCAN algorithm, we developed a set of algorithms which plotted cluster data onto a map using Google Maps API [3] as well as an open source graph generation language called DOT.

Preliminary conclusions for this case study may suggest that tweets coming out of Europe and United States are associated with negative sentiment. In contrast, South Asia, Central Africa and Canada seem to have large clusters associated with positive sentiment.

VII. Conclusion

The purpose of this study was to demonstrate the power of mining unstructured data from an unlikely source of data. We demonstrated that it is possible to make sense of data which carries little to no significance individually by aggregating patterns and using domain knowledge to identify clusters representing similar characteristics. We showed use of advanced text mining techniques to identify semantic meaning as well as morphological similarity in keywords to make clean, standardized data sets that are understandable by a clustering algorithm. We selected healthcare informatics to demonstrate the significance of data for a complex domain. Social networks maintain a loose relationship to most healthcare informatics systems and yet the available data and patterns yielded useful results.

References

[1] Twitter, Inc. “GET Search.” Twitter API Documentation. July 2012. <https://dev.twitter.com/docs/api/1/get/search>.

[2] Russell, M. (2011). Mining the social web. O'Reilly Media.

[3] Google Inc. “Google Maps API”. Google Developers., December 2012. <https://developers.google.com/maps/>.

[4] Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques. (3rd ed.). Waltman, MA: Morgan Kaufmann Publishers.

[5] GeoNames. “GeoNames Web Service Documentation”. December 2012. <http://www.geonames.org/export/web-services.html>.