Cross-Platform Social Media Data Modeling

Deeksha Srinath
Earlham College
Richmond, IN
deeksha.srinath@gmail.com

ABSTRACT

Across the board, pervasive use of social media is generating unprecedented amounts of social data today. Social media provides a wide variety of accessible platforms for users to share information, and connect with others. Mining this data gives businesses, users, and consumers the opportunity to extract patterns. However, social media data is characterised by it's lack of structure, noise and vastness [4]. Thus, the task of mining it proves challenging. The above listed advantages have been attractive enough that we have developed tools, methods and algorithms that allow us to effectively mine social media platforms. With the expansion of different platforms however, there is a need to mine data not only within each platform, but across different platforms. This work focuses on this cross-platform aspect of social media mining. It focuses on developing a cross-platform data model that allows for data from multiple (2) platforms.

Keywords
Social Media, Mining, Data Modeling

1. INTRODUCTION

Data mining research independent of social media has produced numerous methods, tools, and algorithms to handle huge volumes of data. This toolkit of methods and tools allows us to solve real world problems by harnessing the power of data. This type of data mining has become pivotal in fields such as bio-informatics, data warehousing, business intelligence and predictive analytics [2].

The spread of social media and it's resulting ubiquitous presence in our lives allows for the creation of vast amounts of user- generated data that naturally lends itself to the advantages of data mining. Mining social media can help us solve a number of issues ranging from detecting implicit or hidden groups in a social networking site to protecting user security. At it's core, social media mining is an emerging multidisciplinary area where researchers of different backgrounds make important contributions to gleaning important relationships from social data. Much of this work however, has remained platform specific. Increasingly, with the accessibility of web Application Programming Interfaces(APIs), social media mining research is developing as multidisciplinary but equally platform-specific field.

Social media data is characterised as being noisy and unstructured. Removing this nois-e and giving structure to the data is essential before performing effective mining. Social media data are distributed because there is no central authority that maintains data from all social media sites [2]. Removing this noise and providing this structure to a fraction of user data for two social media sites (Twitter, Facebook) is the bulk of the work of this paper. The structure proposed is a unified data model that is able to effectively represent user data from both these sites.

1.1 Data Mining

Data mining is a process of discovering useful or actionable knowledge in large-scale data. Data mining also means knowledge discovery from data (KDD), which describes the typical process of extracting useful information from raw data. KDD can be described as a process that typically consists of data preprocessing, data mining, and postprocessing. These steps need not be discrete and can be combined together based on the needs of the project.

This project will focus on the data preprocessing aspect of the data mining process. The project divides the data preprocessing aspect of data mining into three parts: data access, data modeling and storage, and data cleaning. With the completion of these three parts of data preprocessing ,the data will be ready for analytics. The data produced will be standardised, clean, and in a consistent format. [2].

1.2 Social Media

Social media is defined as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of user-generated content[2].

Social media gives users an easy-to-use way to communicate with each other on an unprecedented scale and at rates unseen in traditional media. These rates of interaction and the popularity of social media has resulted in masses of social data. Each platform is different, and targets different users and facilitates different types of online interactions and communications. This project will be using data sets from two popular social media sites, Twitter and Facebook.

1.3 Mining Social Media

Data generated on social media sites is very different compared to traditionally used attribute-value data utilised in classical data mining. Social media data is vast, noisy, distributed across platforms and unstructured. For example, Facebook and Twitter report Web traffic data from approximately 149 million and 90 million unique U.S. visitors per month.The distinct and proprietary nature of data from such sites causes their data to be distributed as there no central authority that maintains data from all social media sites. This distributed and difference in data formats pose a daunting task for researchers to understand the information flows on the social media.

This project addresses this issue of distributed, non-standard formats of data from social media sites. It focuses on developing a structure that will meaningfully combine user data from two popular sites.

2. RESEARCH COMPONENTS AND CHALLENGES

Using the above understanding of data mining and social media, this work has split the research process into 3 distinct sections: (1) Data Access, (2) Data Modeling and Storage and (3) Data Cleaning

2.1 Data Access

In general, access to extensive social media data is tightly restricted and commercial access is expensive, although this is changing with the advent of Application Programming Interfaces (APIs). APIs ensure that social media data repositories are available through programmable HTTP-based access. This is however, causing the problem of 'siloed' data: data that is inherently isolated, making it difficult to combine with other data sources [1]. This problem of isolated data is exactly what this research is trying to address.

Traditionally, social media data has been available through three avenues:

- Freely Available and Paid for Tools: Examples includes Gnip and Datasift, which allows users to pay for commercial access to Twitter data-sets [1].
- Data Access via APIs: Many social media platforms allow programmable HTTP-based access to a select amount of data via APIs. Examples include Facebook, Twitter and Wikipedia. While this option allows users easy access to a potion of data from their sites, researchers are concerned about the lack of quality of data through these APIs. One of the major concerns about API data and it's use in academic research is it's include limited clarity with regards to the bias in the sample collected and opaque collection processes[5]
- Publicly Available Databases: This includes repositories that can be freely downloaded, such as the Wiki repository, or the Amazon repository. The Terms of Service of large social media sites like Twitter and Facebook however, do not allow these repositories to make large data-sets freely available.

Access to large data-sets from these social media sites is one of the biggest challenges with social media data analysis. This is especially true when one considers the expense of access to this data. These sites make large scale access to their data available for hefty prices. Large corporations typically are able to comfortably afford these prices, and are thus able to gain access to, and therefore leverage the potential of this data for commercial purposes. Academics and researchers have a harder time getting access to the same data since their research interests might not align with the commercial sensibilities of these social media sites.

2.2 Data Modeling and Storage

Representing different social media data meaningfully in a common model to create a holistic data model is the bulk of the work in this portion of the project. The data modeling process implemented in this project was three-fold:

- Entity-Relationship(ER) Diagram Modeling: In order to create a unified data model, this step involved separately charting ER diagrams for both Twitter and Facebook. An ER diagram is a graphical representation of an information system that shows the relationship between objects, concepts and events within that system. This includes charting and representing dependencies between the data, and understanding the flow of the data for each platform. Once this step was complete for both platforms, the next step was to meaningfully trim relevant fields from both structures into a combined model. This created new data dependencies and the data flow for this new unified model is depicted in Figure 1.
- Iteratively testing unified data model with applicable data-sets from Facebook and Twitter: Once the subset of relevant data was chosen and appropriately trimmed, the data model was implemented into a PostgreSQL database on Earlham's cluster[3].
- Cleaning and trimming data: In order to fit the flows and dependencies of the new model, user data from Facebook and Twitter were individually trimmed and cleaned. The cleaning process for these two data-sets were distinct, since their starting data dependencies and flows were platform- dependant. Initially, small fractions of these data-sets were iteratively honed to ensure the logical flow of data in the new model. These

fractions were the first to be trimmed and cleaned according to the requirements of the new models, and set the standard for the subsequent cleaning and trimming of the larger Facebook and Twitter data-sets.

2.3 Data Cleaning

The unstructured nature of social media textual data can be very noisy (i.e., dirty). Cleaning this data is an important part of the research process. Neglecting the cleaning of social media data can cause skewed relationship and patterns that might not reflect the data. Before cleaning the data, familiarity with the common ways social media data can be noisy is useful[1]:

- Missing data: When data existed, but for whatever reason, did not make it to the raw data being cleaned. Problems occur with: a) numeric data when blank or a missing value is erroneously substituted by zero whi-ch is then taken (for example) as the current price.
- Incorrect data: when a piece of information is incorrectly specified such as decimal errors in numeric data or wrong word in textual data.
- Inconsistent data: when a piece of information is inconsistently specified. For example, with numeric data, this might be using a mixture of formats for dates: 2012/10/14, 14/10/2012 or 10/14/2012.

3. ENTITY RELATIONSHIP MODEL AND DATA FLOW

As mentioned in the previous section, an important section of data modeling is the creation of entity-relationship diagrams. Figure 1 represents the final ER diagram for this project. Figure 1 contains four entities: group, comment, post and like. Each of these entities represents a table in a PostgreSQL database. The tables are arranged so that every post is made in a group. Every post has the capacity to contain comments and reactions. The structure of Twitter data does not include groups, therefore all entries in the database from Twitter are listed under one group. Facebook however, allows multiple groups as an organising principle. This is reflected in the data-set, as Facebook data belongs to two different groups. The data-set therefore currently has three unique groups.

The group ID or gid is common to all tables, therefore each post, comment and reaction is associated with a group. Post ID is common to the post, comment and like tables. So, the post ID links every comment and reaction to the original post. Comment ID is common between the comment and like tables, since the comment ID links each comment it's corresponding reactions.

This data-set allows for a range of querying. It includes information about the date and time of each post, along with indications of popularity of the posts through likes. This is of course, in addition to including the entire message in the body of a tweet or a Facebook post.

It is also easily adaptable within reason. Changes to the features of both Tweets and Facebook posts can be easily incorporated into the model. As long as Tweets and Facebook posts are equipped with a unique post id, all the other features of are able to be adapted into this model. This ensures that minor updates to Twitter's and Facebook's policy does not affect the data model dratically.

Facebook allows users to respond to a post with a variety of emotions such as 'like', 'angry', etc. These varied reactions are reflected in the data set. Twitter only allows users to respond to a post with 'like'.

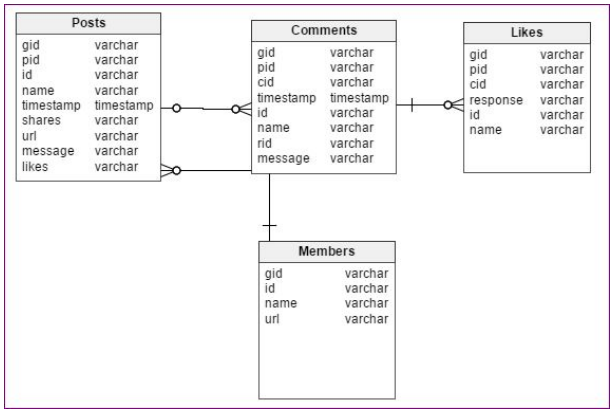

Figure 1: Representation of unified data model for Twitter and Facebook.

The notations on the diagram represent the nature of relationships between the tables. Each member can make multiple posts, and every post can accommodate multiple comments and reactions. This is reflected in data model. For example, as the diagram shows, every post can have multiple comments, but a comment can only have one original post.

4. CONCLUSION AND FUTURE WORK

A size-able amount of the research challenges outlined in this paper comes out of corporations restricting access to their data, or monetizing what they cannot restrict. In addition to monetising data sources, corporations are actively working to isolate their data by making it so proprietary that working with different data sources is extremely challenging.

This is causing a further divide within academia between academics at large universities with vast computational resources and budgets, and academics at smaller institutions interested in the same work. It is important that as researchers we continue to fight for access to big social media data for experimentation. Analysis of social media data can help us with more than targeted marketing and suggested friends lists. It can, and has been, helping solve wider issues such as disaster preparation and response. In order to harness the true cross-disciplinary, cross-platform power of social media data, we need to continue to emphasise it's societal importance through projects that use the power of data to do more than targeted advertisements.

One of the reasons social media data access and challenges associated with it are prevalent today is the lack of incentivizing the free availability of high-quality, search-

able data-sets in consistent formats is part of this process. Institutional processes such as the peer-review process and the tenure process do not currently prioritise and incentivise the creation, preservation and upkeep of standardised data-sets. This is especially true of "new" forms of data such as social data. Changing this attitude towards the availability of data resources centered around social media is crucial to making social media research more accessible.

Future work involves building a web application that allows users to query the unified data model on patterns of eating disorder incidences across Facebook and Twitter.

5. ACKNOWLEDGMENTS

Charlie Peck provided constant and much needed direction and perspective on the purpose and utility of this work. Xunfei Jiang led and co-ordinated the Earlham Computer Science Senior Capstone for Fall 2016. Many future computer science majors and faculty provided much needed feedback and encouragement on the usability and importance of this work. Craig Earley in particular, and the Computer Science senior class in general, provided weekly, iterative and useful feedback. To all of these people, thank you. This work would have lacked much of it's guidance and inspiration without you.

6. REFERENCES

[1] B. Batrinca and P. C. Treleaven. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1):89–116, 2015.

[2] P. Gundecha and H. Liu. Mining social media: a brief introduction. *Tutorials in Operations Research*, 1(4):1–17, 2012.

[3] N. Jatana, S. Puri, M. Ahuja, I. Kathuria, and D. Gosain. A survey and comparison of relational and non-relational database. *International Journal of Engineering Research & Technology*, 1(6), 2012.

[4] A. K. Munk, M. S. Abildgaard, A. Birkbak, and M. K. Petersen. (re-) appropriating instagram for social research: Three methods for studying obesogenic environments. In *Proceedings of the 7th 2016 International Conference on Social Media & Society*, page 19. ACM, 2016.

[5] K. Weller and K. Kinder-Kurlanda. Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research. *Standards and Practices in Large-Scale Social Media Research: Papers from the*, pages 28–37, 2015.