Data Modeling for Cross-Platform Social Media Data: Building a Unified Structure and Dataset

Deeksha Srinath

Fall 2016, CS 488 Senior Capstone

Today, use of social media is generating unprecedented amounts of social data. Mining this data is allowing businesses, users, and consumers the opportunity to extract useful patterns.. Increasing proliferation of social media platforms makes mining across multiple platforms relevant and useful. This project aimed at creating a unified data model that is able to house both Facebook and Twitter data, and creating a clean, unified,, query ready dataset across these platforms.

1 DATA ACCESS:

facebook	twitter
Two-way friendships	One way following
Common Threads	Lists
Comment Likes	Retweets
Disconnect	Mentions
Groups	Block Users

Figure 1: Differences in Twitter and Facebook data

Big data access divides:
Social media data analytics is charecterised by an industry/academic divide. The academic realm is further divided into: (1) who have access to vast computational resources and funding to buy data access (2) those who have limited computational resources and rely on freely available data sources.

Acknowledgments:

-Charlie Peck, Project Advisor,
-Xunfei Jiang, Capstone Insstructor,
CS 488 Fall 2016 Senior Capstone Class,
Jose Ignacio

2 DATA MODELING:

An Entity Relationship diagram(ERD) is a graphical representation of an information system that shows the relationship between objects, concepts and events within that system. Constructing a unified ERD for Facebook and Twitter data involved:
1) individually identifying, defining and choosing entities from both platforms
2) Determining interactions between the chosen entities and determining their cardinality
3) Implementing the diagram into a functional model

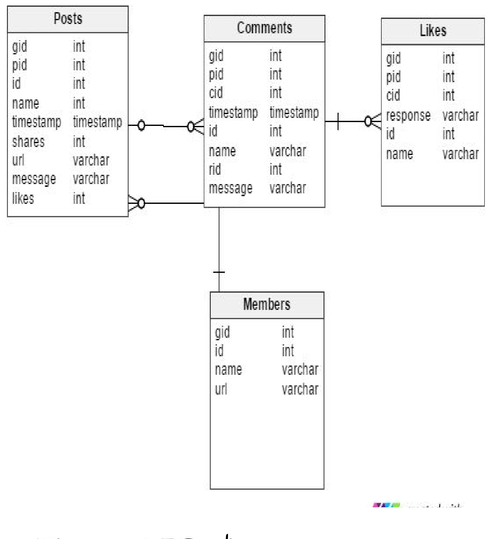

Figure 2:ER diagram representing unified Facebook and Twitter data model

3 DATA CLEANUP:

1) Missing data:
When data existed, but did not make it to the raw data being cleaned..
2) Incorrect data:
When pieces of data are wrongly specified,.
3) Inconsistent data:
When data occurs in non-standard formats.

4 IMPORTANCE AND RELEVANCE:

Part of the motivation to start this work was lack of current institutional incentive.. Even in an academic setting, none of the dominant research processes such as the peer-review process, the tenure and promotion process, or the research process itself currently prioritise the production and integration of high quality, searchable and reusable datasets.

5 FUTURE WORK:

Future planned applications of this structure and model include Web Applications directed at studying eating disorder patterns and observable incidences across Twitter and Facebook,