Twitter Analytics: A Big Data Management Perspective

Published: January 16, 2014

By: Oshini Goonetilleke, Timos Sellis, Xuizhen Zhang, and Saket Sathe

Reviewed by Oscar Solorzano and Kyle Matyac

Background

- Twitter is a massive source of data
  - Opinion mining
  - Event detection
  - Spread of pandemics
  - Celebrity engagement
  - Analysis of political discourse
- Over 200 million monthly active users producing 500 million tweets

Problem Definition

- Many different tools exist for Twitter data analytics
  - Data Collection
  - Data Management Frameworks
  - Languages for querying Tweets
- There isn’t a unified framework for Twitter data management.

Technical Highlights of Problem Solving

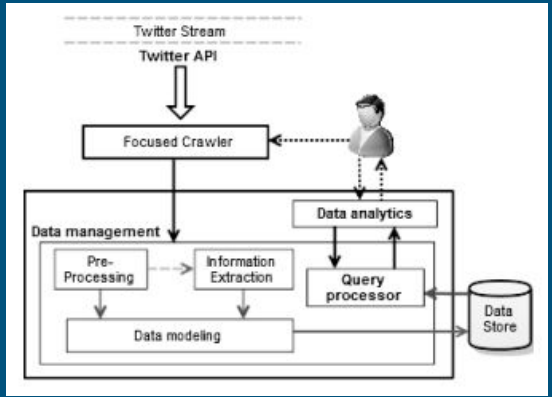
- GeoScope finds trends based on correlations of location-topic pairs.
- TwitterZombie captures hierarchical relationships in tweets.
- TwitHoard takes user keywords to search for hashtags and creates a model.

Table 1: Overview of related approaches in data management frameworks.

	Prepossessing	Examples of extracted information	Social and/or other interactions captured?	Data Store
TwitterEcho [16]	✓	Language	Yes	Not given
Byun *et al.* [19]		Location	Yes	Relational
Twitter Zombie [15]	✓		Yes	Relational
TwitHoard [69]	✓		Yes	Graph DB
CoalMine [72]			No	Files
TrendMiner [61]	✓✓	Location, Sentiment, NEs	No	Key-value pairs
TwitIE [17]	✓✓	Language, Location, NEs	No	Not given
ESA [76]	✓	Location, NEs	No	Not given
Baldwin *et al.* [11]	✓✓	Language, Location	No	Flat files

Technical Highlights of Problem Solving

- These tools support different parts of workflow
- There is a need for an integrated solution
- Unified framework components
  - Focused crawler
  - Pre-processor
  - Data Model
  - Query Language



Our Opinion/Review

- Effectively lays out main components to a unified system that need to be addressed.
- Shows why current components and frameworks for Twitter data analytics work well but are ultimately insufficient for a generic solution.

References

http://kdd.org/exploration_files/16-1-2014.pdf

Project Proposal:

Analysis of Electricity Rates Based on Location, Company, and Zone

Oscar Solorzano and Kyle Matyac

Background/Motivations

- While the current data is very well organized, it lacks any capability to find correlations between its data types.
- It would also benefit from visualizations of the data.
- It is also incredibly long, with tons of repetition since it has one focus being the zip codes.

	A	B	C	D	E	F	G	H	I
1	zip	eiaid	utility_name	state	service_type	ownership	comm_rate	ind_rate	res_rate
2	35218	195	Alabama Power Co	AL	Bundled	Investor Owned	0.105761195	0.060292437	0.114943267
3	35219	195	Alabama Power Co	AL	Bundled	Investor Owned	0.105761195	0.060292437	0.114943267
4	35214	195	Alabama Power Co	AL	Bundled	Investor Owned	0.105761195	0.060292437	0.114943267
5	35215	195	Alabama Power Co	AL	Bundled	Investor Owned	0.105761195	0.060292437	0.114943267
6	35216	195	Alabama Power Co	AL	Bundled	Investor Owned	0.105761195	0.060292437	0.114943267
7	35210	195	Alabama Power Co	AL	Bundled	Investor Owned	0.105761195	0.060292437	0.114943267
8	35211	195	Alabama Power Co	AL	Bundled	Investor Owned	0.105761195	0.060292437	0.114943267
9	35212	195	Alabama Power Co	AL	Bundled	Investor Owned	0.105761195	0.060292437	0.114943267

Objective

- The data warehouse will hold simplified information on electricity rates.
  - It will take a CSV file and convert it to MySQL.
  - Based on star-schema.
- The data mining will be used to answer a few questions:
  - Are there areas with significantly high/low rates?
  - Are there companies that constantly have higher rates than others?
  - Are some zone types favored over others in some areas or by some companies?

Methods (in progress)

- Conversion of CSV to MySQL will be done in Java, unless we find a suitable tool to convert it for us.
- We will likely use a tool to mine; further understanding of the data will be required to choose what algorithms to use.
  - We will likely look for segmentation algorithms for spatial comparisons and association algorithms for comparisons.
- Visualizations:
  - Spatial data will be shown over a google map.
  - Comparison data will probably be in the form of graphs.

Schedule

What to Complete:	When:	Week:
Data Warehouse	April 8	10
Front End	April 22	12
Data Mining/Visualizations	May 6	14

References

http://catalog.data.gov/dataset/u-s-electric-utility-companies-and-rates-look-up-by-zipcode-feb-2011-57a7c

http://www.eia.gov/electricity/data/eia861/