Mining Text and Social Streams: A Review

Author: Charu C. Aggarwal

Publication: ACM, December 2013

•••

Team 09:
Jesus Gomez and Jammy Loeur

Background and Problem Definition

Text mining is the process of obtaining valuable information from text sources such as online articles, social posts, and messages

- Social networks, news aggregator services, & web crawlers all produce text data
- Valuable for gauging current user interest, and presenting relevant information to users

Problems with text data

- The massive volume of data which must be processed
- Lack of structure
- Constantly changing, or evolving

Methods Introduced

Clustering

- Online Spherical k-Means Algorithm (OSKM)
  - Divides the incoming stream into small segments, each of which can be processed effectively in main memory
- Micro-Clustering
  - Creates summaries from the data points which are used to estimate the assignment of incoming data points to clusters

Classification

- One-Class Classification
  - Uses training data of only one specific class of streams to label new streams
- Rule-Based Expert System
  - Uses the position of the words in generating the classification rules
- Neural networks
  - Incremental update process with a network of perceptrons and corresponding weights associated with term-class pairs to classify text streams

What We Learned

Analyzing text data is a relatively new concept that will see growth and further development as the technological era continues

Text streams are an important resource

- Can be used to measure user interest (marketing)
- Present current and relative information to users

Room for improvement in current mining strategies

- Being able to effectively and efficiently analyze massive text streams in real time (social media)
- Language independent text mining

Project Proposals

Data Warehousing & Data Mining

Team 09:
Jesus Gomez and Jammy Loeur

Proposed Projects and Motivations

DW

- Homicide Reports, 1980 - 2014
  - Data of over 22,000 homicides from the FBI's Supplementary Homicide Report.
- Motivation
  - Contains interesting data that we'll be able to optimize and clean.

DM

- Default of Credit Card Clients
  - Dataset showcasing customer information for a Taiwanese bank, as well as default payments if applicable.
- Motivation
  - What correlation does education, age, or income play in the likelihood that customers will default.

Proposal Objectives and Methods

DW

- Objective
  - To clean errors in dataset and transform textual data into numerical for efficient data mining.
- Methods
  - Perform data cleaning and create star schema.

DM

- Objective
  - To analyze the connection among available attributes and customer default on next payment.
- Benefits
  - Better understand characteristics that lead to default payments. Risk management.
- Methods
  - Naive Bayesian

Strategy for Success

General

- Communication is key
- Know the material / dataset
- Set a schedule and have backup plans

Schedule

- Progress Report: April 1
  - Start the DW and DM projects.
- Final DW and & DM Reports: May 2
  - Finish before due date to allow for unexpected circumstances.
- Oral Presentation: May 2
  - Create and prepare a great slides for the presentation, again before deadline.

References

Aggarwal, Charu C. "Mining text and social streams." ACM SIGKDD Explorations Newsletter 15.2 (2013): 9-19. Web. 19 Mar. 2017. <http://www.kdd.org/explorations/P5>.

Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Waltham, MA: Morgan Kaufmann, 2012. Print.

Jyzaguirre. "U.S. Homicide Reports, 1980-2014 | Kaggle." Kaggle. N.p., n.d. Web. 18 Mar. 2017. <https://www.kaggle.com/jyzaguirre/us-homicide-reports>.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.