Name: Subbu Kandhaswamy | IST-707 | HomeWork1|10/9/2019

**Task 1:** Review data mining concepts and tasks Answer the exercise questions 1-3 in Textbook 1.7.

1. Discuss whether each of the following activities is a data mining task or not.

Data Mining Tasks:

* Predicting the future stock price of a company using historical records. This is part of Predictive modeling.
* Monitoring seismic waves for earthquake activities. This is a data mining task, and this will be part of Classification
* Monitoring the Heart Rate of Patience for abnormalities for detecting anomalies and might fall under Classification as well depending on scenarios (abnormal or normal)

Non-Data Mining Tasks

* Dividing the customers of a company according to their gender - Querying
* Dividing the customers of a company according to their profitability – Accounting
* Computing the total sales of a company- Also Accounting
* Sorting a student database based on student identiﬁcation numbers. – Querying
* Predicting the outcomes of tossing a (fair) pair of dice. Mathematical calculation can achieve
* Extracting the frequencies of a sound wave.

1. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving speciﬁc examples of how techniques, such as clustering, classiﬁcation, association rule mining, and anomaly detection can be applied. (Eg., Microsoft e-Store)

It can improve Microsoft to boost their marketing and productivity aligned to meet requirements and raise profits.

Clustering - would help to group results with related or similar themes (Top Xbox games types for Adventure, Racing, Simulator, etc., )

Classification – To help recommend the pre-defined items depending on categories (Application Software, Devices, Gears etc.,).

Sequential association analysis – helps with the purchasing behavior of customers and can detect the probability of customer buying patterns.

Anomaly detection can help MS to alert unexpected or unusual events or behaviors in the product sales. It gives clues where to look for problems and often helps with answers to weirdness in sales.

**Task 2:** Practice your critical thinking and writing

Read the following two news articles. One criticized Google Flu Trend, and the other defended it. Write one paragraph to summarize the criticism, and another paragraph for the defense. Write the third paragraph to offer your own thought, e.g. is the criticism valid? Does the defense make sense? What other problems or benefit do you see in Google Flu Trend or similar big data applications?

<http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/>  
<http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/>

**Criticism:** Once the poster child for the power of Big Data Analytics, Google flu trends was criticized big-time for widely overestimating the number of flu cases in the United States of America. A few months after their announcement, Google flu completely missed the swine flu caused by H1N1 influenza, which was big failure credibility was questioned. And it consistently over-estimated the flu cases for two consecutive years by a large percentage of 50. The google estimates were high in 100 out of 108 weeks which indicates how poor the accuracy of these trends was. After this setback, people were skeptical of Google’s Flu trend algorithm which was used for prediction. The comparative value of the algorithm as a stand-alone flu monitor was questionable. Also, after the update of the algorithm, the google flu trend overestimated the cases by 30%. The recent trend of CDC reports from Doctors on influenza-like illness which lag by two weeks was a more accurate predictor than Google Flu trends.

**Defense:** Google Flu trend considered 40-flu related queries to predict the prevalence of the flu. It worked well in 2008and could predict flu trends for 2009. There was a lot of criticism after its overestimated cases, but the main goal was to complement traditional practices. It was observed that when Google Flu Trends when combined with CDC’s standard monitoring, a better result than either could provide were obtained. Greater value can be obtained wit Google flu trends with other real-time health data. Google flu trend was a system to assist traditional surveillance system and not replace them. Later, when a team examined how to build a better influenza model, the google flu trend model was helpful. It was the only source of external information to provide statistically significant forecast improvements over the base model. Though Google Flu Trend did not live up to its expectations, it has become a base model to develop anything that has predictive ability. Researchers have found GFT and its methods to be useful and relevant.

**My Thoughts:** Criticism that Google flu trend overestimated the cases and it was a failure are valid. But looking at the broader perspective, GFT has laid a foundation for Big Data Analytics. The defense makes sense. Tracking 45 flu-related terms over billions of searches, monitoring trends and, making correlations was a huge win for a big data approach. GFT has set an example for further big data analysis. The system served its purpose of complimenting the traditional surveillance system. The problems like over-fitting, finding the right signal in the noise, data integration, speed, and scalability are found in big data applications. Whereas, big data applications also set a base for the future generation, help to identify business levers and measuring priorities.