**IST 707**

**HW 1**

**MAYA MILEVA**

**Task 1: review data mining concepts and tasks**

1. **Discuss whether or not each of the following activities is a data mining task.**
2. Dividing the customers of a company according to their gender

*No, because is a trivial task. Not new knowledge. We already know their gender, we are not trying to find it, we just sort it. It can be answered with database query.*

1. Dividing the customers of a company according to their profitability.

*No, we already know their profitability. It can be answered with a simple filter in excel or a database query if the probability is an attribute. If profitability of each customer is one of the attributes in customer records, using a threshold can divide the customers according to their profitability.*

1. Computing the total sales of a company.

*No, we already know the sales, we can just perform the simple operation in excel.*

1. Sorting a student database based on student identification numbers.

*No, the answer is not new knowledge, task can also be done in excel, it is a simple database algorithm.*

1. Predicting the outcomes of a tossing a (fair) pair of dice.

*No, this activity is not a data mining task because predicting the outcome of tossing a fair pair of dice is a probability calculation, which doesn't have to deal with large amount of data or use complicate calculations or techniques.*

1. Predicting the future stock price of a company using historical records.

*Yes, this is a data mining task. One could use predictive modeling and use regression to predict the future stock prices based on historical records. Use regression since this is a continuous target variable*

1. Monitoring the hearth rate of patient for abnormalities.

*Yes, this is a data mining task. It is called anomaly detection. By observing the heart rate of the patient, this data mining task can identify the abnormalities if the characteristics of the heart rate are different from normal observations.*

1. Monitoring seismic waves for earthquake activities.

*Yes, this is a data mining task. Anomaly detection task is helpful to identify seismic waves.*

1. Extracting the frequencies of a sound wave.

*No, this is not a data mining task since there is no prediction or description of data.*

1. **Supposed that you are employed as a data mining consultant for 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.**

Data Mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses or other information repositories. There are various data mining functionalities and each of these can be applied in order to improve the company’s search engine.

Clustering - *is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields.* In the context of search engine, clustering can help to display the results that not only contain the keyword specified in the search, but also related results. If we search for ‘wallpaper’ we will get results not only for the word ‘wall’ but it can also display the one with key words ‘design’ or ‘tail’ etc..

Classification - *is the process of finding a set of functions that describe and distinguish data classes or concepts and using this functions to predict the class of object whose class label is unknown. Classification analyzes class-labeled data objects whereas clustering analyzes data objects without consulting a known class label.* This is more of an internal implementation. A list of research papers associated with a keyword could be provided by the search engine. This is done by using either classification rules or decision tree or any other classification algorithms on a set of data whose list of research papers are known and then applying that function to the keyword.

Association rule - *is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.* Search engine can append additional information in its results based on keywords entered by the user. For example, if you a looking to buy a washer you might also be interested in a dryer. Returning results for both washer and dryer keep the search engine one step ahead(identifying web pages that are accessed together).

Anomaly detection - *Anomalies are the data objects that do not conform to the general behavior of the data. The analysis of anomalies is known as anomaly detection.* In cases such as fraud detection, an anomaly is more important than the rest of the data. A search engine can use anomaly detection to avoid displaying results that are not relevant to the searched keyword. Search engine is going to remove all the irrelevant to the key word search results using anomaly detection.

1. **For each of the following data sets, explain whether or not data privacy is an important issue.**

*Data privacy, also called information privacy, is the aspect of information technology (IT) that deals with the ability an organization or individual has to determine what data in a computer system can be shared with third parties.*

1. Census data collected from 1900-1950.

*No, because the information about the people is masked and can not lead to identification of the individuals. In addition census data collected during the years of 1900-1950 is too old and should not pose as a data privacy threat.*

1. IP addresses and visit times of Web users who visit your Website.

*Yes, IP address can be used to find user`s location and identifying him and his personal information. Also all the cookies that are add to your IP every time you visit different website, so data privacy here is an important issue.*

1. Images from Earth-orbiting satellites.

*Yes, to an extent. Depends on how detailed an image could be. For example, if these images fall into the wrong hands, they can easily find out the location of an important person and jeopardize the security of that person.*

1. Names and addresses of people from the phone book.

*Yes, if somebody know your first and last name can easily get your phone number. Nowadays there are so many scams on the phone, people pretending to be family in trouble and asking elderly people for money.*

1. Names and email addresses collected from the Web.

It depends, junk mail and viruses get sent through mails jeopardizing financial accounts and privacy. So, it is a data privacy issue in this case.

**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?

*Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of Flu and Dengue fever based on search patterns. The historic estimates produced by Google Flu Trends and Google Dengue Trends are available below. It is still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue – we're excited to see what comes next. Academic research groups interested in working with us should fill out this form.*

*Sincerely,*

*The Google Flu and Dengue Trends Team.*

**Google Flu Trends (GFT)** was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. The idea behind Google Flu Trends (GFT) was that, by monitoring millions of users’ health tracking behaviors online, the large number of Google search queries gathered can be analyzed to reveal if there is the presence of flu-like illness in a population. Google Flu Trends compared these findings to a historic baseline level of influenza activity for its corresponding region and then reports the activity level as either minimal, low, moderate, high, or intense. These estimates have been generally consistent with conventional surveillance data collected by health agencies, both nationally and regionally. This project was first launched in 2008 by Google.org. The goal was also to build a model that would complement other models in providing accurate predictions. GFT was eventually phased out due to a lot of criticism saying that more than 90% of the predictions were inaccurate. But, did it really deserve all the criticism? Did it achieve the goals it`s supposed to achieve? The following summaries on two of the many articles on GFT give an idea on what was perceived and what went wrong with GFT.

**Google Flu Trends - the limit of big data (Criticism)**

***“****Google Flu Trends, once a poster child for the power of big-data analysis, seems to be under attack.”* said Steve Lohr in his article criticizing the power of google and the limitation of big data. The article states that GFT wildly overestimated the number of flu cases in the United States. It also paraphrases another article “The Parable of Google Flu: Traps in Big Data Analysis,” and talks about the skepticism of the GFT’s algorithms.

A follow-up analysis by four respectful authors (Alessandro Vespignani, David Lazer, Rayan Kennedy and Gary King)found that simply using the trend of C.D.C reports from doctors on influentza-like illness, would have been more accurate predictor than GFT. Their technical criticism was that GFT was not using a broader array of data analysis tools. Indeed, their analysis showed that combining Google Flu Trends with C.D.C data, and applying a few techniques works best. The co-inventor of GFT agreed with that, it was never meant to be a forecasting tool, more like “complimentary signal”. GFT was designed to give a warning signal two to three weeks ahead of the CDC reports. Steve Lohr also cites the papers written by the two computational social scientists Ryan Kennedy and Gary King where they write about the data reversal occurrence where companies have more data than academics and hence the need for researchers to be both technical and have knowledge about the evolution of sociotechnical systems.

**In Defense of GFT (Defense)**

The article defending the new analytical model, which attempted to predict the prevalence of the flu from searches that users made was written by Alexic C. Madrigal. He is writing about how C.D.C had been involved in shaping how GFT functioned. The new toll was celebrated in the news media (CNN, The New York Times, the Wall Street Journal and many more). Flu trends even fit the golden image of Google, circa 2008. Google’s image became even more bigger and shone brighter resulting in a new section being formed called google.org under a guy named Larry Brilliant who envisioned kids in countries like Africa or Cambodia to be able to find out about disease outbreaks faster. Once talks about ‘big data’ emerged and privacy became a bigger concern, google shut down google.org and started focusing on its core thing.

David Lazer pointed to problems in the journal Science, with the opacity of Google`s methods and the inconsistency of the google search user interface and algorithms. The response to the article were rough. The author exclaims how people missed an important fact written in the same paper that GFT predictions will work well when you combine it with CDC model. The author notes how GFT should be considered as a success when compared with the parameters laid out by the *Nature* paper and how it was misunderstood and miscredited by some people whom he terms as ‘Big Data Acolytes’. All the headliners could of made you believe that GFT had gone terribly wrong.

Matt Mohenni speared the creation of Google Flu Trends with Jeramy Ginsberg. At the beginning the were using their (now all but eliminated) “20-percent of the time” to see if they can measure the disease incidence from search query data. After they got sign-off they contacted CDC. Mohebbi pointed that “The goal was to build a complimentary signal to others” – the creator simply didn`t wanted to enmesh their data with the CDC`s because then it couldn`t act as a separate was of understanding a given epidemical scenario. The other problem incorporating with CDC data was that public health officials were most interested in when flu trends deviate from a simple “project ahead a few weeks” kind of model.

In 2013 John Hopkins published a research how to build a better influenza model. The Flu Trend data “was the only source of external information to provide statistically significant forecast improvement over the base model.” It seems like everybody was missing the fact that the role of the GFT was simply to provide complimentary signal. But there was another question: why the model wasn`t working as good as it did in its initial runs. Perhaps the model overfitted flu data or perhaps Google adding more suggestions for users threw off the baseline, or underlying media ecosystems or user search behavior were changing more quickly than anticipated. The author also gets Matt’s view on the criticism who says: "There's huge promise with these techniques, but you have to understand how they should be used." The author concludes the article by saying that the hype surrounding the technology was false hope and that’s why it was declared as a failure. But, this shouldn’t matter and the only thing that should matter is how researchers have found GFT useful and relevant.

**Conclusion**

People have always something to say. I think that the criticism of the model was overexaggerated – it`s the same today, people want a scandal, a big story. The goal of the model was simply forgotten. Magazines or Journals shouldn’t have jumped into conclusions without proper investigation or understanding of big data. Matt Johebbi and other participants from Google were clear with what they are going to build and how its going to help or complement other models. Hence, it’s unfair to say that GFT did not perform well or to term it as a failure. There are anomalies with GFT that could be reduced or eradicated by combining it with other models. This is something the CDC could’ve explored into. Anyone who is familiar with the Big Data arena would know that constant improvement is needed for any model and no model is perfect. One must constantly improvise and reengineer the predictive models to match with fast growing data.

Some other points to agree with from the article written by Steve Lohr is about the need for researchers to be both technical and have knowledge about public health. Its not just to use the data from companies but also understand how the data-driven algorithms of Google, Facebook, Twitter and many more companies influence what people find out in their search engines.

GFT’s failure doesn’t erase the value of big data. What it does do is highlight a number of problematic practices in its use—what we like to call “big data hubris.” The value of the data held by entities like Google is almost limitless, if used correctly. That means the corporate giants holding these data have a responsibility to use it in the public’s best interest. There was persistent pattern of GFT performing well for two to three years and then failing significantly and requiring substantial revision. The point of was not to bury big data but research has demonstrated the value of big data in modeling disease spread, real time identification of emergencies, and identifying macroeconomic changes ahead of traditional methods. But while Google’s efforts in projecting the flu were well meaning, they were remarkably opaque in terms of method and data-making it dangerous to rely on Google Flu Trends for any decision-making.

Google’s sequel to GFT, done right, could serve as a model for collaboration around big data for the public good. Google is making flu-related search data available to the CDC as well as select research groups. A key question going forward will be whether Google works with these groups to improve the methodology underlying GFT. Future versions should, for example, continually update the fit of the data to flu prevalence—otherwise, the value of the data stream will rapidly decay.