Case Study Analysis Report

Visualisation for Data Analytics (CSC-40048)

1 Introduction

In this report, we are going to select two case studies that use diverse data visualization and analytic tools and techniques to solve real world problems. We will investigate details of case study, TweetViz: Twitter Date Visualization and the case study of Detecting influenza epidemics using search engine query data.

In the first case study, they developed a web tool called TweetViz, for visualizing twitter data which offers variety of visualization services that can monitor user activity as well as the trends in the key searches in twitter. In the second case study, they are attempting to provide an early detection of disease activity, by means of the data obtained by monitoring health-seeking behavior from online web search queries, which are submitted by millions of users

The rest of this report is organized into case evaluation in Section 2, With the details of case studies that we are going to critically evaluate with Conclusion in Section 3.

2 Case Evaluation

2.1 TweetViz - Twitter Date Visualization

The goal of this case study is to create a web tool called TweetViz for analyzing and visualizing data generated from the famous micro-blogging service Twitter and use this data to analyze user interest and behavior, detect trends amongst group of users as well as the general Twitter activity linked to a keyword or hashtag. This case study also presents a novel method for visualizing topic distribution in a set of tweets over a period of time. Here advanced Natural Language Processing (NLP) techniques such as LDA (Latent Dirichlet Allocation) algorithm is deployed to generate topic distributions. Anyone interested in exploring Twitter activity can utilize this web tool, which provides a good visual approach of analyzing data from Twitter.

2.1.1 Critical analysis on Visualisations

TweetViz can be divided into two separate modules. The first is user-centric and focuses on inspecting user behavior and the second module is search term orientated, where a person can explore Twitter activity surrounding a particular hashtag or keyword. Moreover, TweetViz utilizes the LDA algorithm for visual representations of topic distribution, from tweets. One benefit of this tool is that it builds visualizations on updated data, unlike some techniques that use static, beforehand retrieved data.

TweetViz uses a few third-party libraries such as Google Charts and d3(Data Driven Documents) for constructing various kinds of visualizations. For producing topic distributions with LDA, they are using Python gensim framework. Twitter permits third party applications and developers to get access to the tremendous quantity of records generated through users each and every day. This is achieved by using the Twitter REST API which gives a lot of distinct endpoints for retrieving this data. In this web tool, they have retrieved tweets from a single user. They also leveraged the search abilities supplied via the Twitter Search API. This is used to retrieve tweets which comprise a specific search term, both keyword and hashtag. But there are few limitations while using Twitter API, first of all there is a limit for the number of requests that can be sent. Also, the search service provided by way of Twitter does not index all tweets available and as a result, we cannot get the complete relevant data. But we will get a sample of the date for this case analysis.

Preprocessing steps need to be taken to visualize frequent keywords in any kind of text analysis. While processing the natural language, we need to apply tokenization and eliminate all the stop words from the text. In addition, words should be lemmatized and stemmed.

1. **User-orientated visualizations:**

   Here we are focusing on a specific user. To understand what the user is interested in tweeting about and to supply insight into his behavior on Twitter, TweetViz explores exclusive approaches to developing interactive visualizations.

   We could see TweetViz plots charts depicting the number of tweets the user posted daily as well as user activity in various parts of the day. We could observe the trends and changes in the user behavior. Figure 1 and Figure 2 are introduced as a stacked column chart that displays discrete hashtags the user tweeted about over some time interval. This offers an exceptional visual way of seeing what the user is involved in and even realizing what type of matters he tends to be involved in the future.

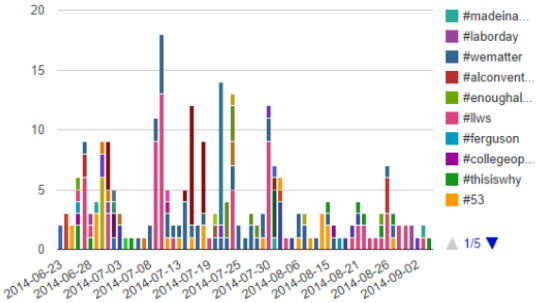

Figure 1: User-hashtag distribution [1].

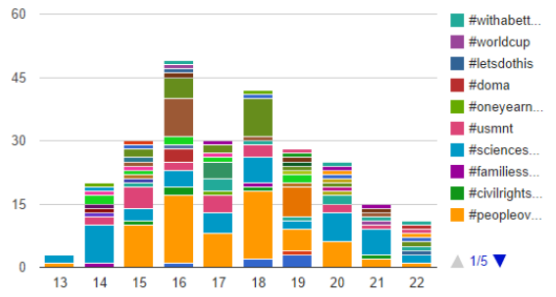

Figure 2: User-hashtag distribution in different time of day [1].

There are some other sorts of visualizations where one can see the influence of the user by means of looking at the wide variety of retweets and favorites his tweets get. This is an easy visualization that can supply facts about the user's impact in the Twitter community by means of a word cloud. In addition to the preprocessing steps described above, Because of the unique domain, the list of stop words wants to be extended with some unique Twitter words and abbreviations such as "RT" ,"retweet", "cc" and so on. The more frequent the word is in a text, the heavier its weight is in the cloud, So the remaining part is weighting process, where extra frequently used words get larger dimensions in the word cloud as hostile to less frequent ones. This is a quality way of staring at what a user tweets about that are no longer targeted to hashtags only. Figure 3 depicts the word cloud.

Figure 3: Word cloud [1].

2. **Keyword-orientated visualizations:**

   TweetViz offers users the opportunity to visualize Twitter activity by searching for a precise hashtag or any given keyword. Users can view a chart displaying the variety of tweets sent containing the search term and its associated timestamp. By doing this we can monitor the spikes and trends in Twitter activity related to that topic. Figure 4 depicts the popularity of a hashtag in exceptional instances of the day. The word cloud visualization is a better alternative because when a user enters a keyword, or a hashtag contributes to a better perception of the context of the search term.

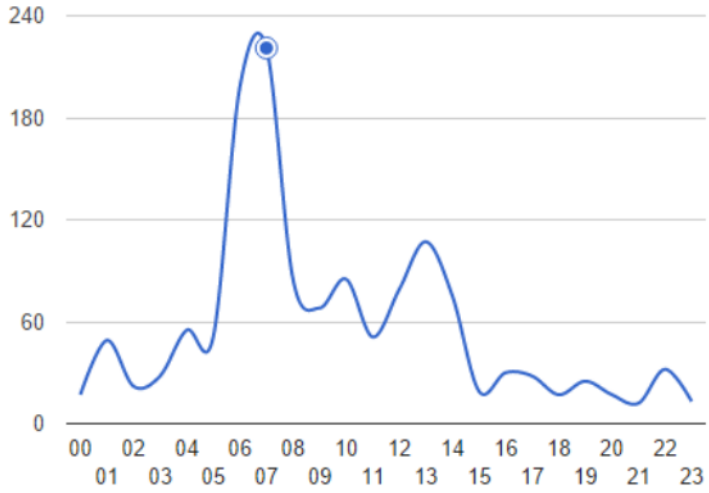

Figure 4: Temporal distribution of a keyword or hashtag [1].

3. **Visualizing topic distribution:**

   In this case study, they have used an advanced NLP algorithm, LDA, in which every topic is modeled as a limitless combination over an underlying set of topic probabilities. By correlating with our data, we could say a tweet is represented as a set of topics associated with appropriate probabilities, and every topic is made up of words with respective probability distribution. When producing the models, we need to perform the preprocessing steps as described above. This topic distribution model can provide insight into how user interests change over time. A good technique to visualize topic distribution in a time interval is to use a Streamgraph, resemble a river-like stream. This makes Stream Graphs aesthetically pleasing and more engaging to look at rather than looking at complex stacked graphs.

   A Streamgraph consists of a finite range of layers, each layer represents a time series. In our case, each layer highlights a topic, and we can track user interest in the subject with respect to the time interval. In this case study implementation of a Streamgraph done in $d3^2$ and silhouette algorithm is used for generating it. They have produced a decent Streamgraph with distinct color schemes on individual layers. One of the major limitations of the LDA algorithm is the fact that we need to predefine the topic count. For this case study they have selected topic count as 20 and displayed 10 subjects or layers in the Streamgraph for simple visualization experience for the users.

   Twitter records that are on the Streamgraph are separated into time slices. Each time slice consists of a set of tweets. As a result, time slices containing extra tweets will have large Y axis values. If we further investigate it, we could observe that the topics with larger differences in distribution to the top and bottom of the Streamgraph as opposed to those with lower variations that cease up in the middle. This adds to a clearer way of providing the layers and differentiating between them. Users are introduced with the phrases that the topic consists of and their respective probabilities. The words show up in a stemmed form, however still,

they are informative and can be used to recognize what the layer is representing. Figure 5 illustrates the corresponding Streamgraph.

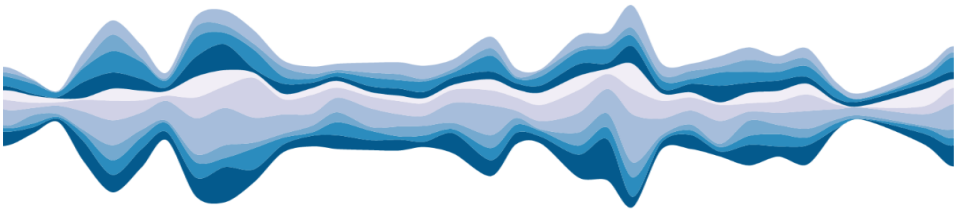

Figure 5: Streamgraph displaying topic distribution [1].

2.1.2 Recommendation

Overall, we could see that the mentioned visualization approaches succeeded in solving the case study objective. That is to detect user interest and behavior as well as the general Twitter activity linked to a keyword or hashtag. We could observe it is not only limited to that objective but can be used to predict the best marketing strategies too. For example, from the User-orientated visualizations, we came across temporal analysis and found the best time to tweet. By identifying the best possible time, we could post our tweets to get maximum reach and impressions.

Since TweetViz tool already uses advanced NLP algorithm, LDA to visualize topic matter distribution, we could improve to a next level visualization via performing sentimental analysis. Then we can monitor and visualize sentiment in tweets about specific topics. So, we could track the user activity from a sentimental perspective too, like how happy the users are.

2.2 Detecting influenza epidemics using search engine query data

The goal of this case study is to present a model of analyzing large numbers of Google search queries to track influenza-like illness in a population. By using this approach, we could see relative frequency of search queries related to the influenza epidemics is highly correlated to the percentage of physician visits in which a patient displays influenza-like symptoms. Using this information, we can precisely estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day, whereas Traditional surveillance system such as US Centers for Disease Control and Prevention (CDC), rely on both virological and clinical data, consisting of influenza-like illness (ILI) health practitioner visits and publishes national and regional data normally with a 1–2-week reporting lag. So, this method may additionally make it possible to use search queries to discover influenza epidemics in a rapid way in areas with a massive population of web search users.

2.2.1 Critical analysis on Model

Proposed model in the case study used an early version of google trends, which is an automated method of discovering influenza-related search queries which shows what people are searching for on google in a particular region; to measure where and when people had the flu. In this study they are evaluating historical logs of online web search queries submitted between 2003 and 2008 across nine regions in United States with a large population of web search users. In this case, they did not preprocess the search query, instead they have used the complete search query with exact sequence of terms issued by a Google search user. That is, they did not apply linguistic variations, synonyms, cross-language translations, misspellings, or sub-sequences in the search query. This can lead to lesser accuracy for the model. By using the IP address associated with each search query, we could detect the general physical location which aids in better analysis and can be used to prevent the outbreak of influenza-epidemic or pandemic situation from the specific geolocation.

They created a simple model that estimates the probability that a random physician visit in a specific region is related to an influenza-like illness (ILI); this is equivalent to the percentage of ILI-related physician visits. The probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below, was employed as

an explanatory variable. Using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query, they have fitted a linear model of the following:

$$logit(P) = \beta 0 + \beta 1 * logit(Q) + \epsilon$$

Where P is the percentage of ILI physician visits, Q is the ILI-related query fraction, $\beta 0$ is the intercept, $\beta 1$ is the multiplicative coefficient, and ϵ is the error term. logit(P) is the natural log of P/(1-P) [2].

Their models were built using publicly available historical data from the CDC's US Influenza Sentinel Provider Surveillance Network (http://www.cdc.gov/flu/ weekly) and they devised an automated technique for identifying ILI-related search queries that did not require any prior knowledge about influenza. The model best fit the CDC ILI data in each region by using only one explanatory variable, Q. This method was used to test each of the 50 million candidate searches in our database, to discover the search queries that could most effectively simulate the CDC ILI visit percentage in each location. The automatic query selection procedure generated a list of the top-scoring search queries, ranked by mean Z-transformed correlation across the nine regions that we considered. We investigated different sets of n top-scoring queries to determine which queries would be included in the ILI-related query fraction, Q(t). The performance of these models was evaluated using the sum of the queries in each set, and we chose n to obtain the best fit against out-of-sample ILI data across the nine regions, depicted in Figure 6. To sum up, the top 45 search queries yielded the best results for predicting out-of-sample points during cross-validation. The best fit was found by combining the n = 45 highest-scoring queries. Although these 45 search terms were chosen at random, they were consistently associated to ILIs, as seen in the table in Figure 7.

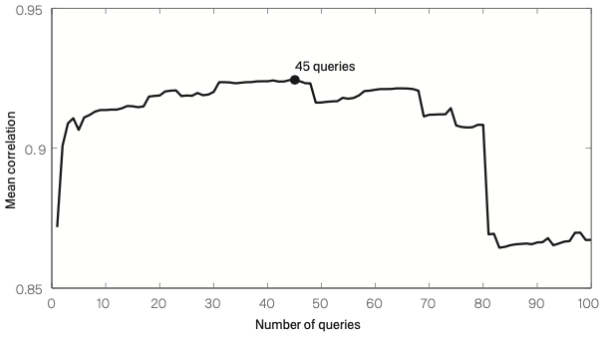

Figure 6: An evaluation of how many top-scoring queries to include in the ILI-related query fraction [2].

	Top 45 Queries		Next 55 Queries	
Search Query Topic	N	Weighted	N	Weighted
Influenza Complication	11	18.15	5	3.40
Cold/Flu Remedy	8	5.05	6	5.03
General Influenza Symptoms	5	2.60	1	0.07
Term for Influenza	4	3.74	6	0.30
Specific Influenza Symptom	4	2.54	6	3.74
Symptoms of an Influenza Complication	4	2.21	2	0.92
Antibiotic Medication	3	6.23	3	3.17
General Influenza Remedies	2	0.18	1	0.32
Symptoms of a Related Disease	2	1.66	2	0.77
Antiviral Medication	1	0.39	1	0.74
Related Disease	1	6.66	3	3.77
Unrelated to Influenza	0	0.00	19	28.37
	45	**49.40**	**55**	**50.60**

Figure 7: Table of Topics found in search queries which were found to be most correlated with CDC ILI data [2].

They yielded 128 training points for each region (each week is one data point) for fitting the model data, then employed 42 more weeks of data (18 March 2007 through 11 May 2008) for final validation. They fit a final linear model to weekly ILI percentages for all nine regions between 2003 and 2007, using this ILI-related query fraction as the explanatory variable, resulting in a single, region-independent coefficient. With a mean correlation of 0.90, the model was able to achieve a decent fit with CDC-reported ILI percentages. The region to which the model was fit yielded a correlation of 0.85 over 128 points, whereas final model was validated on 42 points which yielded a correlation of 0.96 (Figure 8).

They used preliminary versions of their model to generate ILI estimates during the 2007–08 influenza season and shared the results with the CDC's Epidemiology and Prevention Branch of the Influenza Division each week to evaluate timeliness and accuracy. Figure 9 depicts the data provided at various periods during the season. We could see they were able to estimate the current ILI percentage consistently across the nine regions 1–2 weeks ahead of the publication of reports from the CDC's US Influenza Sentinel Provider Surveillance Network. In the mid-Atlantic region, we noticed an increasing ILI percentage during week 5; similarly, the model showed on March 3 that the peak ILI percentage had been attained during week 8, with sharp reductions in weeks 9 and 10. ILI data from the CDC later validated both these findings (Figure 9).

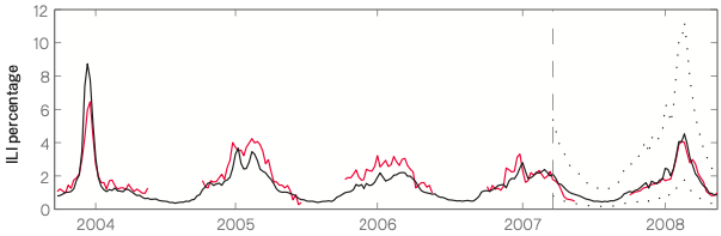

Figure 8: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated [2].

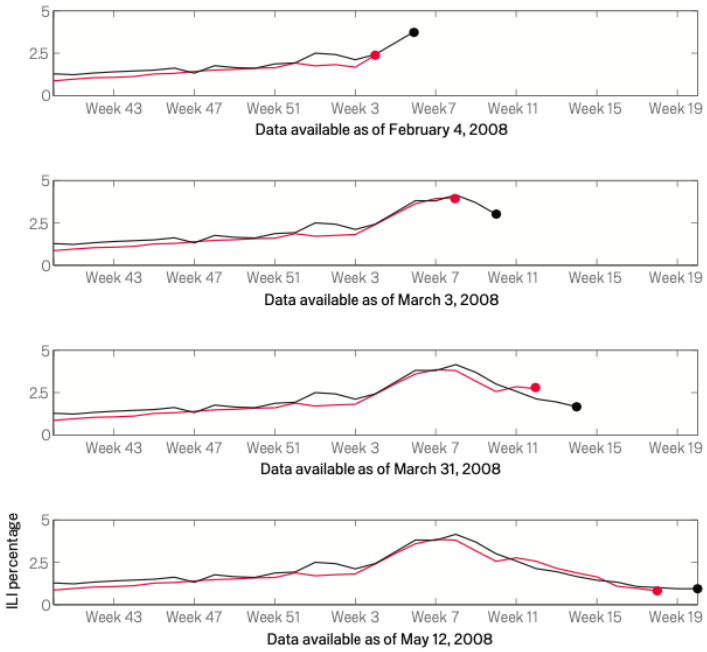

Figure 9: ILI percentages estimated by our model (black) and provided by CDC (red) in the Mid-Atlantic region, showing data available at four points in the 2007-2008 influenza season [2].

By using this model, public health officials and health professionals may be better prepared to respond to seasonal epidemics if they have up-to-date influenza estimates. If a region suffers a rapid increase in ILI physician visits, it may be possible to focus additional resources on that region to determine the outbreak's aetiology, provide additional vaccination capacity, or raise local media awareness as needed. We should also note that this method is not intended to replace traditional surveillance networks because Search queries cannot retrieve demographic data, which is frequently offered through traditional surveillance.

Our model can fail sometimes due to invalid data. For example, Panic and fear among healthy people may lead to an increase in the ILI-related query fraction and exaggerated estimations of the ongoing ILI percentage. Despite strong historical correlations, the model can be prone to false alarms produced by a surge of ILI-related inquiries. Such a false alarm could be caused by an uncommon incident, such as a drug recall for a popular cold or flu cure.

Privacy and Ethics: We should also note that none of the queries in the Google database for this mission can be related to a specific individual. The database retains no records about the identity, internet protocol (IP) address, or precise physical location of any user.

2.2.2 Recommendation

By reviewing similar models, we could see better alternative models for predicting the influenza epidemics using search engine query data. Flu outbreaks can be accurately tracked in real time, allowing public health officials to make timely and significant decisions that could save lives. ARGO (AutoRegression with Google search data) is an influenza monitoring model that employs publicly available online search data. ARGO beats all previously available Google-search–based tracking models, in addition to having a solid statistical foundation. ARGO not only accounts for influenza pandemic seasonality, but also for variations in people's internet search behaviors over time. ARGO is also adaptable, self-correcting, durable, and scalable, making it a potentially effective tool for tracking social events in real time at different temporal and spatial scales. [3].

3 Conclusion

In conclusion, Data analysis can be substantially simplified by first visualizing it, which is more pleasant to the eye and can provide a great deal of information. We have analyzed two case studies and came across various tools and techniques used for data analysis and visualization.

The first case study demonstrates a web tool to analyze and visualize data from Twitter. It is used to comprehend user behavior and interests from a variety of perspectives, as well as general Twitter activity linked to a term or hashtag, and it provides a good visual approach to analyze Twitter data. The second case study created a model which accurately estimates influenza-like illness percentages (ILI) by using Google web search queries. Because search searches can be processed quickly, the ILI estimates that resulted were consistently 1–2 weeks ahead of the CDC's ILI surveillance reports. Early detection afforded by this method could be a crucial line of defense against future influenza epidemics in the United States.

The problem statements were successfully addressed in both case studies. The first case study used dynamic data to illustrate various visualization analyses, whereas the second case study used static historical data to train the model and forecast future trends.

References

[1] D. Stojanovski, I. Dimitrovski, and G. Madjarov, "TweetViz: Twitter data visualization," unknown, Oct. 06, 2014. https://www.researchgate.net/publication/288427532 (accessed Apr. 01, 2022).

[2] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," Nature, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009, doi: 10.1038/nature07634.

[3] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using Google search data via ARGO," Proceedings of the National Academy of Sciences, vol. 112, no. 47, pp. 14473–14478, Nov. 2015, doi: 10.1073/pnas.1515373112.