Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 2, Number 12; July-September, 2015 pp. 45-49

http://www.krishisanskriti.org/acsit.html

# Identification of Epidemic Life Cycle of Flu using Correspondence Analysis

**Amit Verma[1] and Rajeev Mohan Sharma[2]**

*[1,2]Department of Computer Engineering NIT Kurukshetra Kurukshetra, Haryana, India*
*E-mail: [1]amitverma9900@gmail.com, [2]rmsharma123@rediffmail.com*

**Abstract—***These days controlling influenza outbreak is critical issue for health functionaries in any country. It causes thousands of deaths worldwide so that, it must be controlled at an early stage. In this research work, we have done an associated study of algorithms and methods, modeling the outbreak of an epidemic with the focus of swine flu. We have given the significance of the study with respect to micro-blogging website Twitter. We have a given a study from distinctive methods connected to foresee and distinguish flu and considered over the favorable ways in which flu can be detected easily. Based on the limitations of the previous work a novel model has been proposed. An algorithm has been developed for proposed model that takes advantage of single value decomposition and content correspondence analysis for discovering pertinent information that can help us guide to the stages of the epidemic in terms of "Beginning of Epidemic", "Spread of Epidemic" and "Decay of Epidemic". By utilizing this implementation we have the capacity to recognize influenza outbreak more precisely and conquer the confinements of the past work.*

**Keywords***: Correspondence Analysis; Flu Epidemic; Single Value Decomposition; Twitter.*

## 1. INTRODUCTION

Influenza is a virulent disease caused by influenza virus and it is also known as flu. The principal episode of Swine influenza was taken note in 2009. Flu infection causes a respiratory illness of pigs which is called Swine flu, also called H1N1 swine flu, People are generally not contaminated by this infection, but rather here and there, direct exposure to pigs can bring about the transmission. Flu trends are demonstrated by google flu trends or can be anticipated by gathering tweets which are posted on twitter and investigating those tweets. The manifestations of influenza are utilized to anticipate flu trends; numerous individuals with swine influenza have had diarrhea and vomiting. These tweet cases don't clearly tell that it is a swine influenza flare-up [5], however in the event that extensive number of people groups are discussing these side effects and have some connection with the swine influenza is may be anticipated as swine influenza trends. In light of the side effects one can be requested Swine influenza examination, note that a negative result doesn't infer that one doesn't have the influenza. One of the main factor in detection of swine influenza utilizing social networking like twitter is that individuals likewise discuss the swine influenza based test are being directed on them and they examine the results of these outcomes. Neurologic manifestations can be seen in youngsters as a reason for pandemic swine influenza, much the same as regular influenza. These events are rare, accordingly any scientific model made must consider this for forecast of the outbreak, at the same time, according to the seasonal influenza related cases have demonstrated, and these can be serious and even lethal frequently. Trends of seasonal influenza are normally quicker, by and large 1-2 weeks ahead of conventional systems, for example, official reports of CDC [6, 7]. Such data made available by twitter can be utilized to identify influenza trends digitally in time proficient and cost effective way [8, 9]. The digital way of detecting flu trend by utilizing twitter information identifies influenza in its initial stage and it can save a huge number of lives as preventive measures can be taken on time. Early identification of influenza pestilence can help to financial loss misfortune that can be caused by influenza pestilence [10].

## 2. LITERATURE REVIEW

### 2.1 Vasileios Lampos et al. (2010)

In this paper [1], authors said that tracking the expansion of a plague sickness like occasional or epidemic flu is essential undertakings that can lessen its effect. Specifically, early identification and geolocation of an episode are vital parts of this following movement. Different routines are utilized for this checking, for example, tallying the conference rates of general professionals. They used Twitter data. Their strategy is focused around the investigation of a huge number of tweets for every day, looking for side effect-related proclamations, and transforming measurable data into an influenza-score and tried it in the United Kingdom for 2 years throughout the H1N1 influenza epidemic and look at their influenza-score with information from the Health Protection Agency, getting on direct correspondence which is more excellent than 95%. This strategy utilizes totally autonomous information that generally utilized for these reasons, and could be utilized at close time interims, consequently giving reasonable and opportune data about the state of a pandemic. Further

possibilities in area include the use of geographic data and additionally incorporate the mix of other information sources, for instance climate, to enhance the precision of expectations. A summed up variant of this system can likewise be connected to produce naturally the most useful "symptomatic markers", that can permit us to screen more than one pestilence immediately (if their indications are diverse) in different nations autonomously of their dialect.

### 2.2 Harshavardhan Achrekar et al. (2011)

Harshavardhan Achrekar et al. [2] described that Studies have demonstrated that viable mediations might be taken to hold the pestilences if early discovery could be made. Conventional methodology utilized by the Centers for Disease Control and Prevention (CDC) incorporates gathering influenza-like illness (ILI) action information from "sentinel" medicinal practices. ILI report becomes available at the delay of one-two weeks after the patient gets diagnosed. In this paper they proposed the Social Network Enabled Flu Trends (SNEFT) skeleton and uses ARX model which screens tweets with a notice of flu pointers to track and foresee the development and transmission of flu in a populace. Taking into account the information gathered throughout 2009 and 2010, they explored and concluded that the total number of tweets regarding influenza are very much corresponded with the amount of ILI cases as tracked by CDC and further devise auto-relapse models to foresee the ILI action level in a populace. This model anticipates information gathered and distributed by CDC, as the rate of visits to "sentinel" doctors attributable to ILI in progressively weeks. Finally they have concluded that, the real time evaluation, of flu, activities can be obtained through twitter data which can be utilized to predict effectively the current ILI activity levels.

### 2.3 Bumsuk Lee et al.(2012)

In this papers [3] it is elaborated that early detection of flu epidemics and a quick response to that can minimize the impact of the flu. They observed tweets as social signals of flu symptoms to detect the flu epidemics in early stage. They compared a tweet corpus from nine cities in Korea to the weather factors, flu forecast, and Influenza-like Illness datasets. The results show the possibility of using social signals to detect epidemic diseases. They build a system that warns the flu epidemics in real-time using the Streaming API of Twitter. Approach of their research is similar to the Google's flu trends, but used tweets as they are faster to analyze than the analyzing queries on Google.

### 2.4 Jiangmiao Huang et al.(2013)

In this paper [4] they proposed a strategy to find out transmission of contagion by collecting data from Sina Weibo (a micro blogging website in China). In their research work they have collected more than 35.3 million of records over period of 30 moths (August 2009 – Aril 2012) of more than 1 million users. The data set is filtered based on the ILI keywords; only those tweets that contain at least two keywords were considered, which results into 35440 tweets. The keywords used to filter data set are the symptoms of flu such as cough, fever, sore throat, runny nose, muscles aches, headaches, chills, fatigue. The concept of co-location (people living in same city are considered as co-located) and social-tie (people following to a user on social networking site) are considered as two important factors in this transmission model and used Dynamic Bayesian Network to predict the outbreak of flu. The idea they have used to predict the transmission flu is "when infected user moved to other city there is possibility that virus can be spread to his online friend in other cities". In this approach they have used supervised learning to train Dynamic Bayesian Network. The accuracy of this model is greater than 70% if target city is associated with four related cities as observed nodes. Result shows that flu outbreak can be detected by using social sensor data.

## 3. PROPOSED MODEL

In this segment we will be delineated the working of proposed work which likewise endeavors to conquer the limit of the past work done around there. The execution of the proposed structure is portrayed in the accompanying chart shown below:

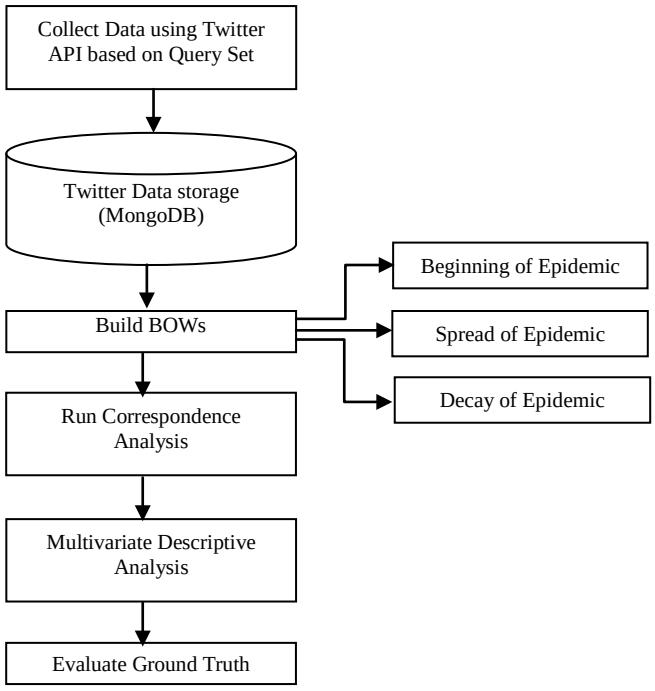


**Fig. 1: Block Diagram of Proposed Model**

### 3.1 Collect Data from Twitter API based on Query Set

Typically, when user tweets, the words used reflect the stage of epidemic spread, these are based on the word/phrase/sentence relevancy score and the stage of epidemic factors that may be accounted for the logic for detection of the epidemic start. Hence, the query set, which

helps to extract relevant tweets and numerical factors helps to calculate the relevance of text are considered as a feature set for doing the research work.

Twitter's prevalence as a data source has prompted the advancement of applications and research in different areas. As Twitter is utilized to stay connected us with our followers and users, we follow and share tweets with each other through the world, and it is also used to give situational attention to an emergency circumstance. Many researchers have utilized Twitter to predict events like earthquakes and distinguish relevant clients to look after to get calamity related data, similarly we shall be doing it to detect the trend of the swine flu outbreak with reference to the some demographics.

A sample of Twitter data can easily be obtained through the APIs which is freely available, to obtain the full view is difficult because the Twitter APIs only allow us to access 1% sample of the Twitter data, which is the result of sampling strategy we will use related to our required information. While collecting Twitter data (runtime) a Query set will be applied, which specifies a set of keywords related to Swine Influenza Activities, so that only useful information stores in MongoDB.

### 3.2 Markov Chain State Model based on BOWs

Once, the collection of Tweets database set is ample with a simple size six months of data, the next step is to build the 'Bag Of Words' for each Epidemic State(Beginning, Spread and Die) through which an epidemic undergoes. So Epidemic state model can be applied on data with respect to BOWs (Bag Of Words), these words, sentences, phrases, verb, adverb, noun verb, pair combinations show the state of affairs in terms of vocabulary tweeted by Twitter handler to its network which will extract the useful content and will divide it into three states as given below:-

- Beginning of Epidemic: This state will indicate the beginning stage of the epidemic.
- Spread of Epidemic: This state will narrate that the epidemic is spreading or has already spread in the specific area.
- Decay of Epidemic: This state indicates that the epidemic in now under control or died.

### 3.3 Correspondence Analysis

Correspondence analysis is an exploratory data analytic technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. As opposed to traditional hypothesis testing designed to verify a priori hypotheses about relations between variables, exploratory data analysis is used to identify systematic relations between variables when there are no (or rather incomplete) a priori expectations as to the nature of those relations.

Correspondence analysis is also a (multivariate) descriptive data analytic technique. Even the most commonly used statistics for simplification of data may not be adequate for description or understanding of the data. Simplification of data provides useful information about the data, but that should not be at the expense of valuable information. Correspondence analysis remarkably simplifies complex data and provides a detailed description of practically every bit of information in the data, yielding a simple, yet exhaustive analysis.

*1) Reduction of Dimensionality*

Another way of looking at correspondence analysis is to consider it as a method for decomposing the overall inertia by identifying a small number of dimensions in which the deviations from the expected values can be represented. This is similar to the goal of factor analysis, where the total variance is decomposed, so as to arrive at a lower - dimensional representation of variables that allows one to reconstruct most of the variance/covariance matrix of variables. However , in the current context , SVD method will be used for doing Correspondence Analysis , in simple words , it *will a way to find only those* tweets that are relevant to Flu tre*nds and reduce the Big Data set*.

*2) Singular Value Decomposition*

By decomposing the total inertia the researcher can identify important sources of information that help describe this association. Using different decompositions will yield different interpretations of the association, and lead to different graphical outputs. The most common type of decomposition used, with a few exceptions, in correspondence analysis is singular value decomposition (SVD). The next subsection describes the use of SVD to perform simple correspondence analysis. Other types of decompositions can be used, and two others are described at the end of this section.

Classically, simple correspondence analysis is conducted by performing Singular Value Decomposition (SVD) on the Pearson ratios. The method of SVD, also referred to as the "Eckart–Young" decomposition, is the most common tool used to decompose the Pearson ratios. For the application to the analysis of contingency tables, Eckart & Young (1936) conjectured that the Pearson ratio may be decomposed.

## 4. RESULTS

In this paper, we discuss the outcomes of the implementation strategy we have built which allows us to achieve the research objectives including the mapping of life cycle of the flu, its detection and it detection using Correspondence Analysis of the Tweets Content. The proposed algorithm is implemented using java and the database used for the saving of the tweets is mongoDB.

Evaluation Parameters: Conventionally, previous researchers have evaluated their work by finding covariance between the Centers for Disease Control and Prevention (CDC) and Influenza Like Illness (ILI) Data. However, in our case, we shall try to find the accuracy of the method to classify the incoming tweet content with respect to the life cycle of the epidemic. Hence, for evaluation, the performance method of evaluation is based on "Recall and Precision" graphs as defined below:

- Average Precision at "k": It is defined as "All retrieved tweets taken into account for particular life cycle stage of the epidemic with respect to the all tweets retrieved for that particular stage in epidemic outbreak" with cutoff / threshold 'k'.
- Average Recall: It is defined as "All retrieved tweets taken into account for particular life cycle stage of the epidemic with respect to the all tweets retrieved for that all stages in epidemic outbreak".
- G - Measure: It is defined square root of the product of recall and precision values. It is helpful when the number of tweets for each life cycle is not equal.
  - $G = \sqrt{(precision * recall)}$ (1)
- Balanced F - Measure: The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.
  - $F1=2 * (precision * recall) / (precision + recall)$ (2)

### 4.1 Graphical Representation of Results

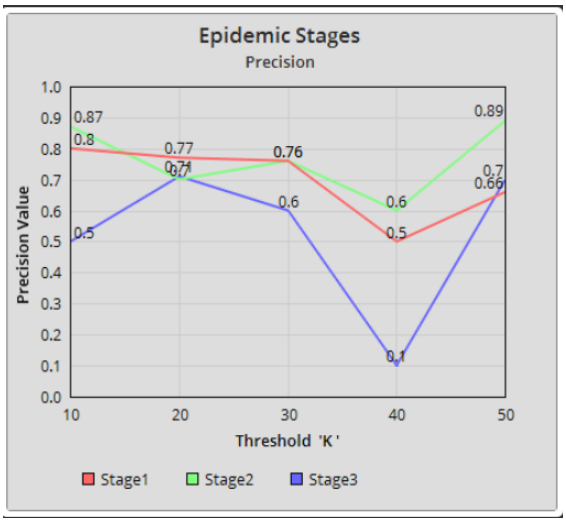


**Fig. 2: Precision Value of All Stages at K threshold**

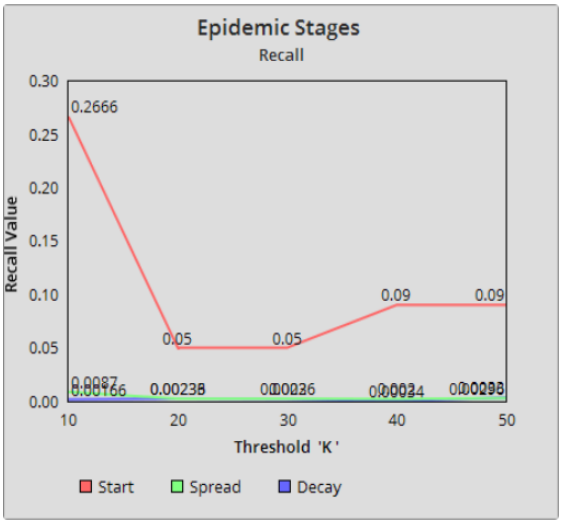


**Fig. 3: Recall Values of All Stages**

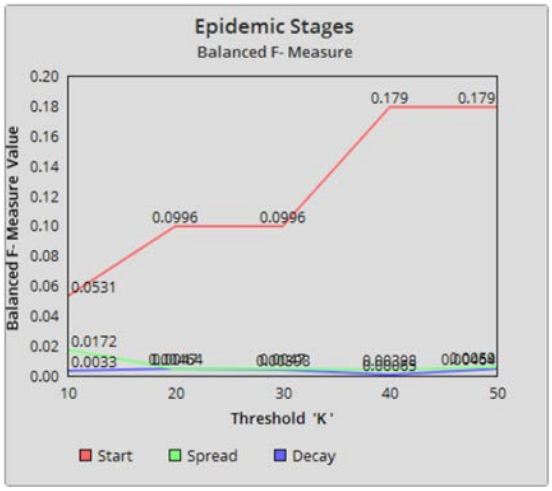


**Fig. 4: F Score Measure of all Stages**

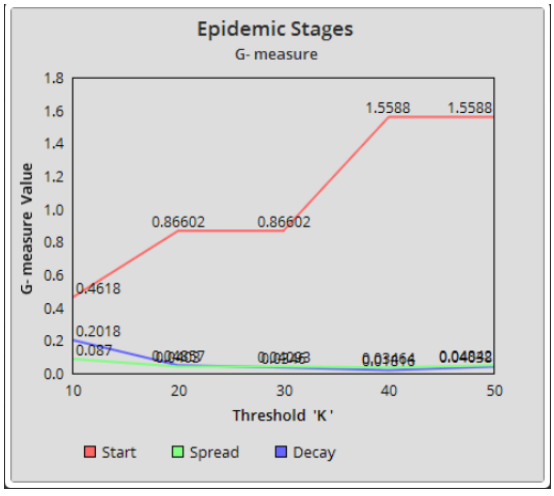


**Fig. 5: Geometric Measure Values of All Stages of Epidemic**

### 4.2 Interpretation of Graphs

It is clear from the graphs that as the precision go up; the recall tends to come down. Usually, precision and recall scores are not discussed in isolation. Instead, either values for one measure are compared for a fixed level at the other measure (e.g. precision at a recall level of 0.75) or both are combined into a single measure. A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional G-measure or balanced F-score. Good precision and reasonable recall for YES would also improve our weighted mixture, because it would improve the F-score for entails and it would also improve overall accuracy.

Since our work was limited from 20 November 2014 to 20 May 2015 and for the fair evaluation of our system random tweets documents were picked from each month to build a corpus of tweet 300 based on these 300 tweets the values of precision and recall were calculated. It is apparent above graphs , the maximum rate of change is occurring in stage 2 which is Spread of Epidemic and slowest rate of change is in stage 3 which is Decay of Epidemic, this is clear from the aggregated values of rate of change which is 3.86 and 0.26.

These results show that there is a rapid change towards 2nd stage, which has an average rate of change calculated on the basis of threshold k. Above all we can infer that dominant epidemic stage in the study period is stage 2. However when we try to calculate the proportion of change which is simply how much of change value a particular change point has contributed to overall change percentage. We find that the 1st stage has the highest average; this is due to fact that between these change points the rate is slower than 2nd stage but little faster than the 3rd stage.

### REFERENCES

[1] V. Lampos, N. Cristianini, "Tracking the flu pandemic by monitoring the social web," 2nd International Workshop on Cognitive Information Processing, pp. 411-416, 2010.

[2] H. Achrekar, A. Gandhe, R. Laszarus, S. H. Yu, B. Liu, "Predicting flu trends using twitter data," Computer Communications Workshops (INFOCOM WKSHPS), IEEE, pp. 702-707, 2011.

[3] B. Lee, J. Yoon, S. Kim, and B. Y. Hwang, "Detecting social signals of flu symptoms,"8th International Conference on Collaborative Computing: Networking, Applications and Worksharing, pp. 544-545, 2012.

[4] J. Huang, H. Zhao, J. Zhang, "Detecting flu transmission by social sensor in china," Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, IEEE, pp. 1242-1247, 2013.

[5] J.Parker, Y. Wie, A. Yates, O. Frieder, N. Goharian, "A framework for detecting public health trends with Twitter," Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, pp. 556-563, 2013.

[6] E. Cho, S. A. Myers, J. Leskovec, “Friendship and mobility: user movement in location-based social networks,” Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1082-1090, 2011.

[7] E. Aramaki, S. Maskawa, and M. Morita, “Twitter catches the flu: detecting influenza epidemics using twitter,” Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 1568–1576, 2011.

[8] C. Corley, A. Mikler, K. Singh, D. Cook, “Monitoring influenza trends through mining social media,” Proceedings of the International Conference on Bioinformatics Computational Biology, ICBCB, pp. 340–346, 2009.

[9] A. Culotta, “Towards detecting influenza epidemics by analyzing twitter messages,” Proceedings of the 1st Workshop on Social Media Analytics, ACM, pp. 115–122, 2010.

[10] V. Lampos, N. Cristianini, “Nowcasting events from the social web with statistical learning,” ACM Transactions on Intelligent Systems and Technology, 2012.