Detecting and Analyzing Influenza Epidemics with Social Media in China*

Fang Zhang[1], Jun Luo[1,2], Chao Li[1], Xin Wang[3], and Zhongying Zhao[1]

[1] Shenzhen Institutes of Advanced Technology
Chinese Academy of Sciences, Shenzhen, China
[2] Huawei Noah's Ark Lab, Hong Kong, China
[3] Department of Geomatics Engineering, University of Calgary, Canada
{fang.zhang,jun.luo,chao.li1,zy.zhao}@siat.ac.cn,
xcwang@ucalgary.ca

Abstract. In recent years, social media has become important and omnipresent for social network and information sharing. Researchers and scientists have begun to mine social media data to predict varieties of social, economic, health and entertainment related real-world phenomena. In this paper, we exhibit how social media data can be used to detect and analyze real-world phenomena with several data mining techniques. Specifically, we use posts from TencentWeibo to detect influenza and analyze influenza trends. We build a support vector machine (SVM) based classifier to classify influenza posts. In addition, we use association rule mining to extract strongly associated features as additional features of posts to overcome the limitation of 140 words for posts. We also use sentimental analysis to classify the reposts without feature and uncommented reposts. The experimental results show that by combining those techniques, we can improve the precision and recall by at least ten percent. Finally, we analyze the spatial and temporal patterns for positive influenza posts and tell when and where influenza epidemic is more likely to occur.

Keywords: Influenza Epidemics, Social Media, Data Mining.

1 Introduction

Influenza is a severe disease and seasonally spreads around the world in epidemics, causing over 3 million yearly cases of severe illness and about 250,000 to 500,000 yearly death[1] Global attention has been drawn to this issue from both medical and technical perspectives. However, influenza is unable to be detected under the traditional surveillance system both effectively and efficiently, thus making the disease monitoring a challenging topic.

In recent years, social media, for instance, Facebook, Twitter, MySpace and TencenWeibo, has become a popular platform among people on which they create, share, and

* This research is partially funded by the NSF of China (Grant No. 11271351 and 61303167) and the Basic Research Program of Shenzhen (Grant No. JCYJ20130401170306838 and JC201105190934A). Xin Wang's research is partially funded by the NSERC Discovery Grant.

[1] http://en.wikipedia.org/wiki/Influenza

V.S. Tseng et al. (Eds.): PAKDD 2014, Part I, LNAI 8443, pp. 90–101, 2014.

propagate information. Social media gains ascendancy over traditional media because of its better performance in stability, fast propagation and efficient resource utilization. Therefore, it is gradually replacing traditional medias and grows in fast pace as the platform of useful information sharing. Recent work has demonstrated that prediction of varieties of phenomena can be made by using social media data. These phenomena include disease transmission [18], movie box-office revenues [4], and even elections [20]. In this paper we illustrate how social media can be used to detect and analyze influenza epidemics in China. Specifically, we consider the task of detecting and analyzing influenza by utilizing the posts from TencentWeibo, one of the most popular social networks with more than 500 million users in China.

We first extract influenza-like posts from our TencentWeibo data corpus. The most common influenza symptoms are chills, fever, runny nose, sore throat, headache, coughing, fatigue and discomfort. Although, these symptoms as keywords can be utilized to determine whether a post is an influenza-like post, inaccurate, ambiguous or keywords related posts might still disturb the collection of the real influenza-like posts such as (all posts and words are translated from Chinese to English in this paper):

- One should have more water when catching flu.
- Avian flu is under epidemics this spring.
- Jesus, fevering, have I got cold?

These posts all mention the word of “flu” or flu symptoms. Nevertheless that does not mean that the posters have been affected by influenza. We consider these posts (news, advices or suspicion) as negative influenza posts. Our goal is to detect positive influenza posts and analysis influenza epidemics in China with TencentWeibo data. As discussed above, it is necessary to extract positive influenza posts from the whole dataset to get more accurate results. In this paper, we propose a machine learning based classifier to filter out negative influenza posts with 0.900 precision and 0.913 recall.

Next, after classification, TencentWeibo data is analyzed and processed from the perspective of time and space respectively. From the perspective of time we can find out which place is more likely for influenza outbreak and from the perspective of space, we can discover when is more likely for influenza outbreak in one city or a certain province in China.

This paper is organized as follows: related works are presented in next section. In Section 3, a short introduction to TencentWeibo and the characteristics of our dataset are provided. In Section 4, several data mining techniques which are used in our research are introduced. In Section 5, evaluation of our model is shown. In Section 6, our model is applied in detecting and analyzing influenza epidemics in China. We conclude and give the future work in Section 7.

2 Related Works

In recent years, scientists have been using social media data or other information to detect influenza epidemics and to provide earlier influenza warnings.

Espino et al. [7] proposed a public health early warning system by utilizing data from telephone triage (TT) which is a public service to give advice to users via telephone in 2003. They obtained TT data from a healthcare call center services and software company. By investigating the relationship between the number of telephone calls and influenza epidemics, then reported a signification correlation.

Magruder [13] utilized the amount of over-the-counter (OTC) drug sales to build a possible early warning indicator of human disease like influenza. Influenza patients requirement for anti-influenza drugs makes this approach reasonable. They reported the magnitude of correlations between clinical data and some OTC sales data and then measured the time lead after controlling for day-of-week effects and some holiday effects.

Ginsberg et al. [8] built a system, utilizing Google web search queries, to generate more comprehensive models for use in influenza surveillance. Their approach demonstrated high precision, obtaining an average correlation of 0.97 with the CDC-observed influenza-like illness (ILI) percentage.

Lampos el al. [11] proposed a regression model, by applying Balasso, the bootstrapped version of Lasso, for tracking the prevalence of ILI in part of UK using the contents of Twitter. Compared to the actual HPA's ILI rates, their model achieved high accuracy.

Aramaki el al. [3] proposed a system to detect influenza epidemics. First, the system extracts influenza related tweets via Twitter API. Next, a support vector machine (SVM) based classifier was used to extract tweets that mention actual influenza patients. Their approach was feasible with 0.89 correlation to the gold standard.

However, these previous approaches ignored some major characteristics of posts that may impede the classification. First, all posts and reposts have 140 word limitation. That could cause limited features we can use in SVM. Second, the reposts could be no comments or the comments without features. We propose words association rules and sentiment analysis to overcome those problems and improve the classification precision and recall.

3 Dataset

3.1 TencentWeibo Dataset

Launched in April, 2010 by Tencent Holding Limited, TencentWeibo is a Chinese micro-blogging (weibo) website, which is extremely popular around China, consisting of more than half a billion of users (0.54 billion users by Dec, 2012[2]). Like Twitter, each user of TencentWeibo has a set of followers, and from this point TencentWeibo can be considered as a social network. Users can upload and share with its followers photos, videos and text within a 140 word limit, known as posts like tweets in Twitter, that typically consist of personal information about the users. The posts composed by one user are displayed on the user's profile page, so that its followers can either just read, comment or repost the same content and post to their own pages. For one user, it is also possible to send a direct message to another user. A repost, called retweet in Twitter, is a post made by one user that is forwarded by another user. Reposts are useful and fast for information spreading, like photos, videos, text and links through TencentWeibo community. Due to its huge amount of users and prevalence, TencentWeibo is

[2] `http://it.21cn.com/itnews/a/2013/0121/11/20247640.shtml`

increasingly used by a number of companies and organizations to advertise products and disseminate information. Mining TencentWeibo data to make the future prediction on some social phenomena has become an innovative approach in China.

3.2 Dataset Characteristics

The data we used in our experiments was obtained by downloading the posts from TencentWeibo.com with TencentWeibo Search API. We used “flu” and its common influenza symptoms, such as fever, runny nose, as keywords to ensure that the posts we obtained were influenza related. We obtained 2.59 million influenza related posts over a period of six months from Nov, 2012 to May, 2013. Most of these posts contain location information which indicates the poster’s living city. By accurately classifying influenza posts from this data set, we can analyze spatial and temporal patterns of influenza epidemics.

3.3 Label Rules

Three annotators are responsible for assigning positive or negative label to every post in both training dataset and test dataset. One post is labeled as a positive only when it meet one of the following requirements.

- Post indicates the poster has influenza. Since each post has one attribute showing the city name which indicates where the post is sent, we can use this information to do spatial analysis of the positive posts.
- The post mentions other person (relative or friend) has influenza and also mentions the location of the other person. For example, one post says “My poor brother got fever in Beijing”. Then we annotate it as positive post and the count of influenza case in Beijing will be increased by one. Otherwise, if there is no indication of location, then the post is annotated as negative post since it has no use to our analysis.
- For reposts, if one repost has no comment, we consider the reposter is consented with the original poster, thus the repost is annotated as the original post’s label. If the repost has comment, we label it according to the previous two rules.

Each of these three annotators individually labeled a post x as negative (-1) or positive (+1) influenza-like post described as y_1, y_2, and y_3. Each post was given the final label by the following function $L = \sum_{i=1}^{3} y_i, y_i \in \{+1, -1\}$, where the positive value of L indicates a positive influenza post, while the negative value of L indicates a negative influenza post.

4 Methodology

We build a support vector machine (SVM) [5, 12, 14, 16] based classifier to classify influenza posts with the help of association rule mining and sentiment analysis. SVMs are well-known supervised learning models used in machine learning, particularly for text classification and regression analysis. In terms of linear SVM, the training data set **D** with points is defined as below.

$$\mathbf{D} = \{(\boldsymbol{x}_i, y_i) | \boldsymbol{x}_i \in \mathbf{R}^{\mathbf{p}}, y_i \in \{+1, -1\}\}_{i=0}^{n} \tag{1}$$

Where $\overrightarrow{x_i}$ is a p-dimensional real vector and y_i is the label of the point $\overrightarrow{x_i}$, indicating to which class $\overrightarrow{x_i}$ belongs. And the classification function of linear SVM is:

$$f(\boldsymbol{x}) = sign(\sum_i \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + b) \tag{2}$$

Where the value of $f(\boldsymbol{x})$ indicates the point's class, α_i is Lagrange multiplier and b is the intercept.

In the rest of this section, several techniques that we utilized in our model will be illustrated.

4.1 Association Rule Mining

Most TencentWeibo posts are short texts with significant characteristics of short length and few features. If potential strongly associated features can be added to the original texts, making longer length and more diversified features, classification performance will be improved. In data mining, association rule learning [1,2,9,21] is a popular and well researched method for discovering interesting relations between variables in large databases. For these reasons, association rule learning is applied to find strong rules in our data.

According to the original definition by Agrawal et al. [1] the problem of association rule mining in our research is defined as: Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n texts features called items. Let $D = \{t_1, t_2, ..., t_n\}$ be a set of posts called the database. In a given database D, an association rule is similar to a form of A⇒B where A, B∈I and A⋂B=∅, the sets of items A and B are respectively called antecedent and consequent of the rule. An easy attempt of differentiation of strong rules is calculating its support and confidence, and thus, mining frequent patterns is the key to obtain strong association rules.

Methods like Apriori [2] can be used for mining association rules and frequency patterns. Apriori is not an efficient algorithm as it needs to find all the candidate itemsets and to repeatedly scan the data base during the process. However, in our research, since frequent patterns with 2 features, such as {cold, runny nose}, are needed and the candidate sets with more than 2 features ($k > 2$) are avoided, Apriori algorithm becomes efficient and is applied in our research.

Frequent patterns with a given minimum thresholds on support and confidence are regarded as strong association rules which then will be utilized to extend short posts to improve classification performance. For example, if "cold"⇒"runny nose" is a strong association rule in our data base, then word token "runny nose" will be added into the texts of the posts which contain word token "cold" as a feature.

4.2 Sentiment Analysis

Sentiment analysis [6, 15, 17, 19] refers to the application of natural language processing (NLP), computational linguistics, and identification plus extraction of subjective information over text analysis[3] Generally speaking, sentiment analysis is designed for

[3] http://en.wikipedia.org/wiki/Sentiment_analysis

acquiring the attitude of the corresponding author or lecturer upon his or her contextual works on a comprehensive level. The basics of sentiment analysis is to categorize the absolute standing or meaning of the given material words based on the opinion delivered into three classification positive, negative, or neutral.

In our research, an important component of sentiment analysis which focuses on the automatic identification for whether a repost contains positive or negative opinion about influenza is to identify the emotion expressed in the posts if the according poster has infected with influenza.

In our dataset, each post that is downloaded by keywords is one of the following three kinds:

- Posts with features after feature selection.
- Commented reposts with an original post, but no features in comment.
- Uncommented reposts with an origin post.

For the first kind of posts, the SVM classifier directly classifies them. As to the rest two kinds, we separate the reposts r into two parts, comment part c and original post part o. Take “Fortunately, I didn’t. || @ someone: I got flu” as an example. “Fortunately, I didn’t. ”is part c and “I got flu.” is part o. After feature selection, however, this post has no features. The SVM randomly classifies it as positive influenza post or negative one; nevertheless, the poster definitely has not got influenza. In this situation, sentiment analysis is needed to improve the SVM’s precision. Our approach can be described as:

$$L(\boldsymbol{r}) = \begin{cases} s(\boldsymbol{c}) \times f(\boldsymbol{o}),\ f(\boldsymbol{o}) = +1 \\ \quad f(\boldsymbol{c}),\ f(\boldsymbol{o}) = -1 \end{cases}, \tag{3}$$

$$s(\boldsymbol{c}) = \begin{cases} +1,\ no\ negative\ word\ in\ c \\ -1,\ has\ negative\ word\ in\ c \end{cases} \tag{4}$$

where $f(\boldsymbol{o})$ is defined in formula 2, $s(\boldsymbol{c})$ indicates whether comment part c has negative attitude to the origin part o (we regard reposters have positive attitude towards the original post if the according reposts have no comment on the original posts), and the value of $L(\boldsymbol{r})$ indicates the repost’s class.

5 Experiments

We collected about 2.59 million posts posted within the time period from Nov 2012 to May 2013, using TencentWeibo Search API. We separated those posts into three groups.

Training data consists of 4092 posts which were randomly selected by computer and annotated by 3 annotators. Then these posts were used for the purpose of SVM classifier training.

Test data consists of 2500 posts randomly selected by computer and annotated by 3 annotators like training data. These data were used to evaluate the SVM based classifier.

Experiment data are the rest of the posts collected. They were used in experiments of influenza epidemics detection and analysis. Those posts were separated into six groups by month within the time range that we studied, from Nov 2012 to April 2013.

Table 1. Examples of positive and negative weighted significant features of our SVM classifier

Positive Weighted Feature	Weight	Negative Weighted Feature	Weight
have cold	1713.39	fatigue	132.99
feel ill	385.70	faint	86.98
fevering	146.85	healthy	56.43
runny nose	85.06	question	47.41
very	82.26	share	46.23
sore throat	48.61	later	43.00
serious	48.22	little	41.64
rhinobyon	40.01	lack	39.99
headache	38.67	sneeze	37.04
seemingly	37.43	nervous	36.01

We first applied feature selection for SVM training because of three reasons: (1) To improve the efficiency of training and testing process. (2) To delete noisy features. (3) To improve classification precision.

We calculated Chi-squared value for every word token that appeared in the training data set. As SVM features, top 1,000 word tokens ranked by Chi-squared value were utilized. Before word segmentation and vectorization, punctuation and special characters were striped, mentions of user names (the“@” tag), reposts (the “||” tag) and expressions (the “/” tag) were removed, and all other language characters were ignored. Table 1 lists examples of significant features we used as SVM features.

Besides precision and recall, $F_1 = \frac{2 * precision * recall}{precision + recall}$ which is a weighted harmonic mean that trades off precision versus recall was utilized to evaluate our classifier.

In our experiment, the kernel of SVMs is linear. The evaluation of this SVM classifier on test set showed 0.79 precision, 0.80 recall and 0.795 F_1. There are two reasons to cause relative low precision and recall:

- A TencentWeibo post consists text with a 140 word limit and most of the posts are short texts with only several words. Not even a single feature is contained in some processed short posts after word segmentation and vectorization. Our SVM classifier’s scheme of random labeling them leads to relative low precision and recall.
- In terms of reposts, reposts without comment are also qualified for the condition above, as reposts with comment always indicate the reposter’s attitude on the original post. However a SVM classifier is not capable of analyzing poster’s sentiment on this kind of post.

We handled the first case by applying rule mining to extend word-segmented posts to obtain more features before vectorization. Based on our training data, we learned some word association rules as shown in Table 2 with the threshold of 0.01 minimum support and 0.6 minimum confidence. However, the performance of the SVM classifier with 0.797 precision, 0.804 recall and 0.800 F_1did not improve too much.

We then utilized sentiment analysis to improve the evaluation of classification. First, we collected thousands of emotional-related words and put them into 2 groups

Table 2. Examples of word associations rules with support and confidence

Feature 1	Feature 2	Support	Confidence
feeling ill	having cold	0.085	0.67
sore throat	having cold	0.011	0.60
runny nose	having cold	0.017	0.65
sneezing	runny nose	0.013	0.65
serious	having cold	0.012	0.69
question	having solved	0.011	0.65
lack	voting	0.011	0.65
bad	having cold	0.015	0.67

Table 3. Examples of both emotional negative and emotional positive words

Negative	no	don't think so	deny	don't agree	disappoint	abhorrent	annoyed	angry	insane	bad
Positive	yes	I think so	accept	agree	gladness	amused	happy	glad	smart	good

(emotional negative and emotional positive). Table 3 lists examples of both emotional negative and emotional positive words. We then applied formula 3 and 4 to classify reposts with an original post.

As shown in Figure 1, when considering word associations and sentiment of posters, the classification performance substantially improves, achieving up to 0.900 precision, 0.913 recall and 0.905 F_1.

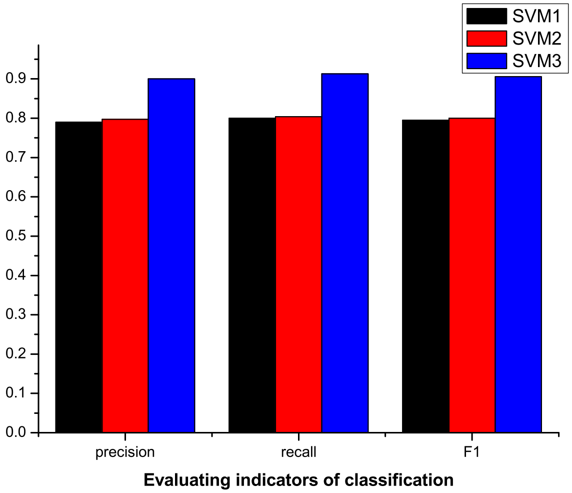
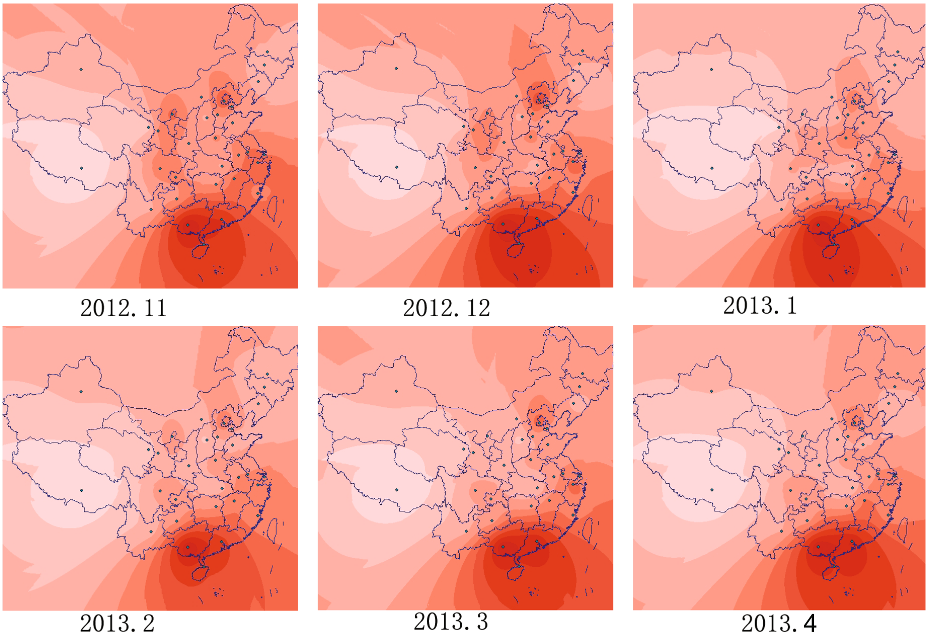


Fig. 1. Summary of evaluation results. SVM1 represents original SVM classifier, SVM2 represents the SVM based classifier with association rule mining, and SVM3 represents the SVM based classifier with association rule mining and sentiment analysis.



2012. 11 2012. 12 2013. 1

2013. 2 2013. 3 2013. 4

Fig. 2. Spatial analysis of influenza epidemics in China from Nov 2012 to Apr 2013. The darker colored area indicates a higher influenza index and the lighter colored means a lower index.

6 Spatial and Temporal Analysis of Influenza Epidemics

A simple model was built on the experiment dataset to monthly estimate the average extent of influenza epidemics of every province in Chinese mainland. For every month from Nov 2012 to Apr 2013, our model automatically calculates the index of each provinces influenza epidemics by summing each provinces positive influenza posts which were extracted from the whole posts corpus with the help of the SVM based classifier, dividing the sum by net citizen scale of the certain province which was obtained from Statistical Report on Internet Development in China published by China Internet Network Information Center (CNNIC) in 2012, and multiplying the result with 10,000 to make the final number more readable. For example, in Nov 2012, in our database Beijing has 6,916 positive influenza posts and 13,790,000 net citizens. Therefore, the influenza index of Beijing in Nov 2012 is $6,916 \times 10,000/13,790,000 = 5.015$.

6.1 Spatial Analysis of Influenza Epidemics

After computing the monthly index of influenza epidemics in each province, we assign that value to the capital city of the corresponding province. The we use Kriging [10]which is a geostatistical estimator that infers the value of a random field at an unobserved location to spatially interpolate influenza epidemics index of the whole map. Figure 2 represents the distribution of influenza epidemics in China from Nov 2012 to April 2013. From Figure 2, we obtain some conclusions as below:

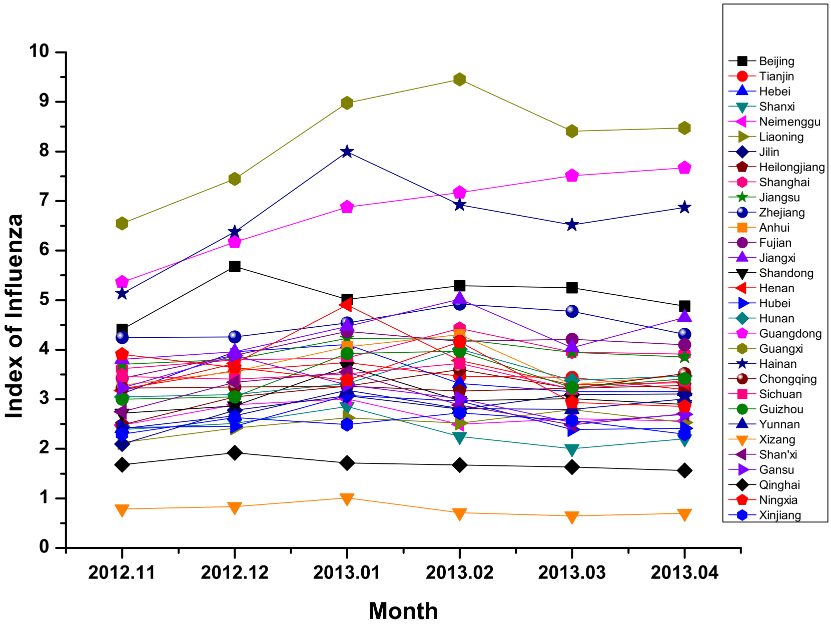


Fig. 3. Time analysis of influenza epidemics in China from Nov 2012 to Apr 2013. Different kinds of dots represent different provinces and each dot shows the influenza index of a given province in the given month.

- With the drop of temperature from Nov 2012 to Dec 2012, high influenza-index regions are increasing and low influenza-index regions are reducing. On the contrary, with the rise of temperature from Dec 2012 to Apr 2013, high influenza-index regions are reducing and low influenza-index regions are increasing.
- From Nov 2012 to Apr 2013, southeast coastal provinces including Guangdong, Guangxi, and Hainan have higher influenza indices. While west provinces including Xinjiang, Qinghai, and Tibet have relatively lower influenza indices.
- Areas such as Beijing-Tianjin, Chengdu-Chongqing, Yangtze River Delta, and Pearl River Delta with bigger fluid population and more density of population have higher influenza indices. Areas such as Xinjiang, Xizang (Tibet), Qinghai, and Gansu where density of population is far smaller than the areas mentioned above have lower influenza indices.

6.2 Temporal Analysis of Influenza Epidemics

Figure 3 represents the influenza indices of each province in China mainland from Nov, 2012 to Apr, 2013. Some conclusions can be obtained by observing Figure 3:

- From Nov 2012 to Apr 2013, Guangxi, Guangdong, and Hainan have relatively higher influenza indices than other provinces. Xizang and Qinghai contrarily have relatively lower influenza indices.

- The influenza indices of Xizang and Qinghai from Nov 2012 to Apr 2013 slightly change. However, the influenza indices of Guangxi, Hainan, and Henan fluctuate a relatively great deal.

7 Conclusions and Future Work

In this paper we propose a TencentWeibo based influenza epidemics detection and analysis model with data mining techniques. Basically, we build a support vector machine (SVM) based classifier to classify influenza posts. In addition, we use association rule mining to enrich the features of posts to overcome the limitation of 140 words for posts. We also use sentimental analysis to classify the reposts without feature and uncommented reposts. Our experimental results show that by combining those techniques, we can improve the precision and recall by at least ten percent. Finally, we analyze the spatial and temporal patterns for positive influenza posts and tell when and where influenza epidemic is more likely to occur.

In future work, we will use more TencentWeibo data to verify our model's efficiency and effectiveness. Also we will focus on the personal prediction of a certain poster whether he or she would catch influenza in the next a few days based on its personal TencentWeibo data.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Record 22, 207–216 (1993)
2. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)
3. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: Detecting influenza epidemics using twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1568–1576. Association for Computational Linguistics (2011)
4. Asur, S., Huberman, B.A.: Predicting the future with social media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 492–499. IEEE (2010)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
6. de Haaff, M.: Sentiment analysis, hard but worth it!, customerthink (2010), http://www.customerthink.com/blog/sentiment_analysis_hard_but_worth_it
7. Espino, J.U., Hogan, W.R., Wagner, M.M.: Telephone triage: a timely data source for surveillance of influenza-like diseases. In: AMIA Annual Symposium Proceedings, vol. 2003, p. 215. American Medical Informatics Association (2003)
8. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature 457(7232), 1012–1014 (2008)
9. Hipp, J., Güntzer, U., Nakhaeizadeh, G.: Algorithms for association rule mining - general survey and comparison. ACM SIGKDD Explorations Newsletter 2(1), 58–64 (2000)
10. Krige, D.G.: A statistical approach to some mine valuation and allied problems on the Witwatersrand. PhD thesis, University of the Witwatersrand (1951)
11. Lampos, V., De Bie, T., Cristianini, N.: Flu detector - tracking epidemics on twitter. In: Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M. (eds.) ECML PKDD 2010, Part III. LNCS, vol. 6323, pp. 599–602. Springer, Heidelberg (2010)

12. Lin, C.-J.: A guide to support vector machines, Department of Computer Science, National Taiwan University (2006)
13. Magruder, S.: Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. Johns Hopkins University APL Technical Digest 24, 349–353 (2003)
14. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
15. Melville, P., Gryc, W., Lawrence, R.D.: Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1275–1284. ACM (2009)
16. Meyer, D., Leisch, F., Hornik, K.: The support vector machine under test. Neurocomputing 55(1), 169–186 (2003)
17. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
18. Sadilek, A., Kautz, H., Silenzio, V.: Predicting disease transmission from geo-tagged microblog data. In: Twenty-Sixth AAAI Conference on Artificial Intelligence, p. 11 (2012)
19. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12), 2544–2558 (2010)
20. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 178–185 (2010)
21. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)