Sina Weibo Incident Monitor and Chinese Disaster Microblogging Classification

Hua Bai [1*], Xunguo Lin [2], Bella Robinsion [2], Robert Power [2]
[1]School of Management, Harbin Institute of Technology
Heilongjiang, 150001, China
[2]CSIRO Digital Productivity Flagship
G.P.O. Box 664, Canberra, ACT 2601, Australia
baihua1727@163.com

JDIM
Journal of Digital Information Management

ABSTRACT: *This paper describes the initial work on developing an all-hazards emergency event detector using messages obtained in near-real-time from the public timeline of the Chinese Sina Weibo microblogging service. The system filters target keywords corresponding to emergency events of earthquakes, floods, typhoons, fires and storms and then uses classifiers to identify messages from people experiencing the corresponding emergency event. Then, this study carried out experiments that compare the performance of four different classification methods and also explore to improve the classifier by the new training data captured by SWIM recently. After Chinese text pre-processing, feature selection and training set size, the experimental results demonstrate Random forests classifier could get best performance but need more long time to run in R, thus the potential to improve this classifier for setting up the SWIM system need to be explored in the future. While similar work has been reported using Twitter content, this is the first time these techniques have been applied to the Sina Weibo microblogging service for multiple emergency event types. This paper also outline the experience of accessing Sina Weibo messages, provide a summary of their structure and content, note the challenges faced in processing this text using Natural Language Processing packages and outline the developed website for users to view the processed messages. The long term aim is to develop a general emergency notification and monitoring system for various disaster event types in China reported by the public on Sina Weibo which can be used by the appropriate emergency services as a source of improved situational awareness.*

Subject Categories and Descriptors
H.2.8 [Database Applications]: Data mining; **K.4.1 [Public Policy Issues]:** Privacy

General Terms: Text Mining , Chinese Text Classify

Keywords: Sina Weibo, Disasters Monitor, Microblogging Classification, SWIM

Received: 19 December 2014, Revised 10 February 2015, Accepted 17 February 2015

1. Introduction

According to the latest annual report on humanitarian crises and assistance from the United Nations Office for the Coordination of Humanitarian Affairs[1], 97 million people were affected worldwide by national disasters in 2013. China was the most affected country with 27.5 million impacted, followed in orders by the Philippines, India, Vietnam and Thailand. The biggest disaster events in China during this time in terms of cost were earthquakes (US$6.8 billion) followed by Typhoon (US$5.7 billion for Japan and China combined). The overall global trend for the cost of disasters has been steadily increasing over the past 10 years[1].

The response and recovery activities to manage disaster events are typically performed by emergency services agencies that are specifically trained to deal with the situation appropriately. Large scale disasters may involve the armed forces and in some countries international aid agencies help also. Coordinating the efforts of these multiple groups to achieve the best outcome in the shortest time frame is a challenging task and central to these activities is effective and accurate information sharing of the impact to the environment, the people affected and infrastructure damage. This is referred to as situational awareness and it is vital that all those involved share a common operating picture.

Social media has been recognized as an emerging new source of information for emergency managers[2,3,4]. Twitter in particular is an important channel of communication to source content from people experiencing disasters and for emergency services agencies to inform the public of what's going on. For example, Olteanu et al.[4] found that on average 12% of Tweets during natural disasters events were from eyewitnesses. After examining a sample of disaster related tweets they found 15% of messages were from affected individuals, 14% were offering caution and advice and 9% noted information about affected infrastructure and utilities. Similarly, research from the American Red Cross[5] found that 28% of American citizens choose social media services to send messages after disaster events and that 20% obtained emergency information from a mobile application. They also found that 40% of citizens would use social media to inform their contacts they were safe if impacted by an emergency event and if they were to send are quest for help via social media, 70% expected help to arrive in less than three hours of posting.

Twitter has been a widely investigated source of crowd sourced emergency event information[6,7,8,9,10]. This service is not available in China and we wanted to explore how well similar techniques reported using Twitter can be used on publicly available messages from a Chinese microblogging platform. Sina Weibo was chosen since this is essentially the Chinese equivalent to Twitter, is the most influential Chinese microblogging service and has more than 156 million active users per month with more than 69 million active users per day[11].

The task is to identify emergency events described by people experiencing them in China from their Sina Weibo messages. Automatic classification of messages plays an important role in identifying relevant messages. This investigation is the first step in developing a general alert and monitoring system for disaster events in China from content published on a microblogging service.

This paper is organized as follows. In section 2 we will review the related works including processing Chinese text and Microblogging classification. In section 3 we will introduce our SWIM system, followed by the experimental evaluations about four classifiers using the new training data captured by SWIM in section 4. The conclusions are given in section 5.

2. Related Works

2.1 Processing Chinese text

The Lack of whitespace between words is the main difficulty to pre-process Chinese text before classification. So we need to do the segmentation work at first. There are a lot of software tools to do automatic word segmentation on Chinese text, including the IK Analyzer, Stanford Word Segmenter and Microsoft's S-MSRSeg. We used ansi-seg tool, an open source java tool; it is based on ICTCLAS system developed by Chinese Academy of Science.

The next step is stop word removal. Automatic stop word extraction on Chinese text is an important reach area resulting in several extraction algorithms[12], however, the auto methods are all hard to implement, hence, many researchers and practitioners choose to use stop word list. In this work, we use an open source Chinese stop word list including 73 symbols, 1,113 Chinese words and 9 numbers.

Another difficulty on processing Chinese microblogging is about the traditional and dialect Chinese text. Since they have the same sentence structure and grammar with simplified Chinese, we can apply the restoration approach, converting the Traditional and dialect text to Simplified Chinese before classification[13].

2.2 Microblogging Classification

Compared to news and the other formal text, microblogs are very short, which could cause the extracted features sparsely. In addition, microblogs are often lack formal expression and standard grammar. People love to use a lot of abbreviations, internet words, acronyms and dialect on microblogs. Thus, microblog classification presents several challenges to text classification process. Classification method of short and sparse Chinese microblog is an active research area recently. For our work, we focus on four classifiers: Support Vector Machine (SVM)[14,15], from the category of discriminative classification methods; Naïve Bayes, one of the generative classifiers; K Nearest Neighbor (KNN)[16], a lazy learning algorithm calculating the closest training examples in the feature space and Random forests[17], an ensemble learning method based on lots of decision tree classifiers. All of them have been demonstrated to perform well in traditional text classification research.

3. Sina Weibo Incident Monitor

Sina Weibo Incident Monitor (SWIM) is a multi-disaster detector based on near-real-time microblog messages from China. The system filters key words related earthquake, flood, fire, typhoon and storm(heavy rain) from Sina Weibo messages available on the public timeline and use SVM as the classifier to determine if the messages containing key words correspond to the people experiencing a disaster. We developed a web site of visualization interface for exploring with the messages generated by our SWIM system.

As Figure1 shows, SWIM interface is constituted by four elements: (1) A China map. Users can zoom-in and -out and can also recenter it to a particular region as desired; (2) A GUI (graphical user interface) below the map to let user define keywords and search timeframe; (3) Keyword counts graph. The blue curve shows the message counts containing the search keyword per 15-minute and the red curve shows the emergency message counts also per 15 minutes. Our classification system (described late) judges if a message is emergency (i.e. related to an event just

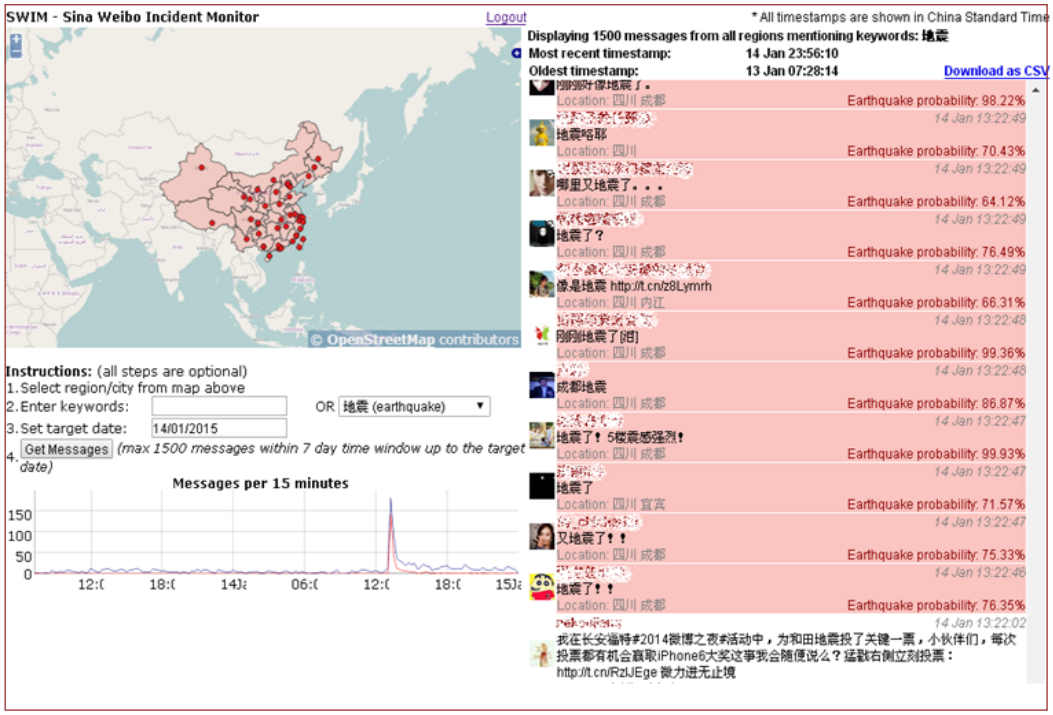

Figure 1. SWIM Web Interface

occurring) or not; (4) Message display area on the right hand side of SWIM. Up to 1,500 microblog messages each contains the keyword are displayed. Messages marked red are those that the classifier believes to be positive and a detection probability estimated by the classifier is also shown.

The aim in this work is to improve our classifier used in SWIM. Since SWIM was developed in September 2014, it has already detected a number of earthquakes occurred in China. However, because the training data (messages) used to build the SWIM classifier was from 2012, and internet language used in microblogging always changes/evolves quickly, we would like to supplement our training data with the latest data to improve the classifier. We also want to explore if the SVM classifier as used by our SWIM system is still the best classifier for our event detection task.

4. Experimental Results

4.1 Training Data

The current training data is made of two parts. The first part comes from our original training dataset (containing 934 messages), and the second part contains 1,913 messages collected by SWIM. There are 12,261 messages containing the keyword "earthquake" between the 27th August and the 3rd December 2014. We manually selected 956 positive (i.e. related to a current earthquake) messages and then randomly picked up 957 negative messages from the pool (12,261 messages).

4.2 Feature Selection

We used the whole training dataset (2,847 messages) to exam the best features combination. We explored eight features including: character count, word count, user mention count, hash tag count, hyperlink count, question mark count, exclamation mark count and unigram. Thus, we conducted a total of 255 experiments for each of the four classifiers (see section 2.2), and ran a 10-fold testing to estimate the Recall (R), F1, Accuracy (A) and Precision (P) for each combination. The best combination of features for each classifier is shown in Table 1.

ClassifierBest	Combination of Features
SVM	Char count, link count, question mark count, exclamation mark count and unigrams
KNN	Exclamation mark count and unigrams
Naive Bayes	Links count, word count, chars count, question mark count and exclamation mark count
Random forests	All features

Table 1. Classifiers' best features combination

4.3 Training Set Size

In order to test if the training dataset for building up the classifier is large enough, we ran an experiment of varying the training data size to explore its effects. We performed ten tests by each classifier using a training dataset of 10% (284 messages) incrementally to reach 100% (2,847 messages). To make sure the results are not affected/biased by a particular sample of data, for each training size we also ran ten times by changing the dataset of the same size. Thus, we processed 100 tests for each classifier to estimate the best data size. The results of each classifier are shown in Figure 2.

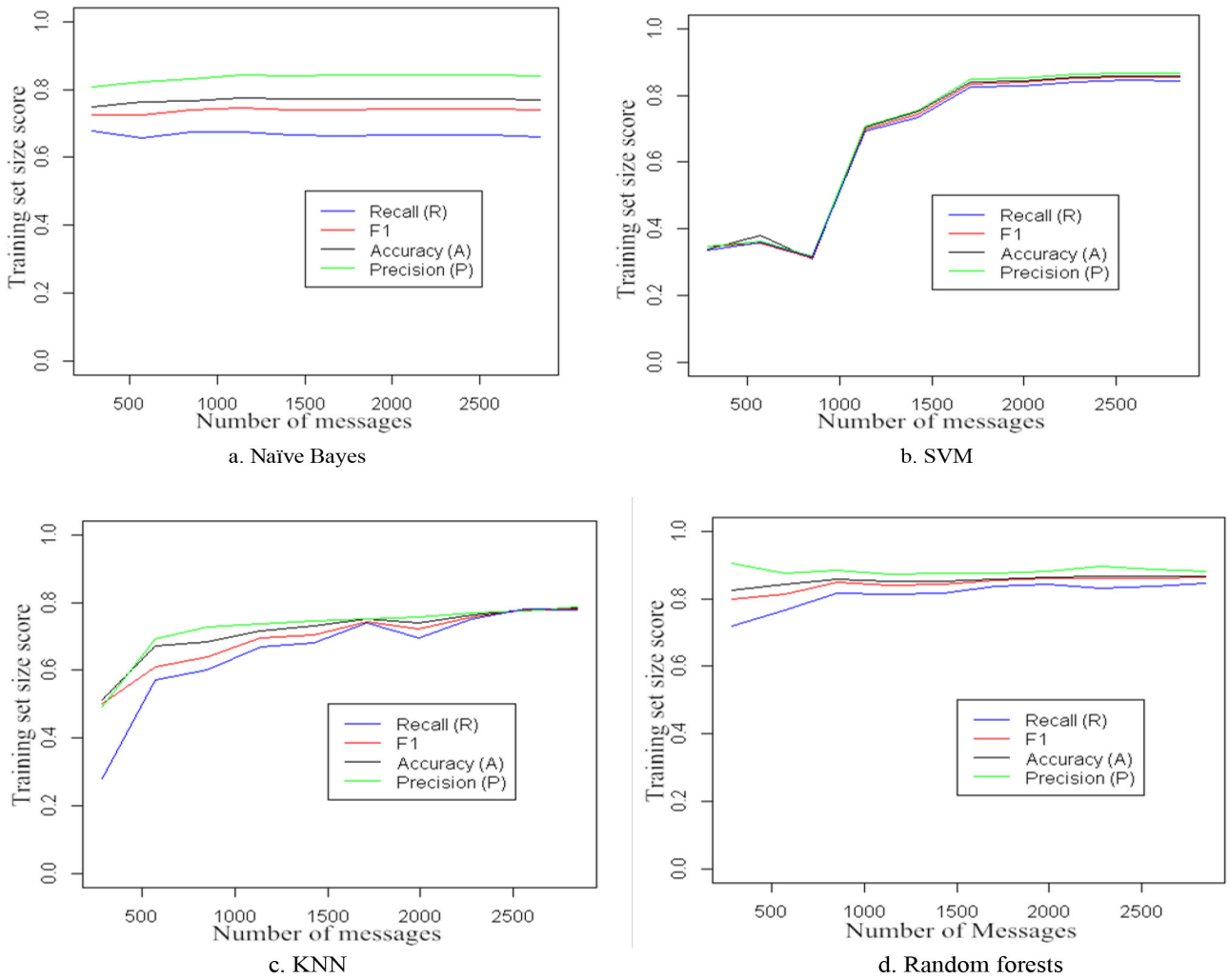

Figure 2. Adjusting the Training Set Size of Four Classifiers

The features used for this experiment were the best combinations identified by feature selection process. From this process we found that Random forests and SVM are better than KNN and Naive Bayes. The Accuracy, F1, Precision, Recall of the best data size for each classifier are listed in Table 2. However, this evaluation was based on our experiment training dataset, which was man-made and balanced. In a real world, microblogging stream is unbalanced and noisy. For further evaluation of the classifiers' performance, we adopted the Receiver Operating Characteristic (ROC) Curve method, which is more suitable for unbalanced data in a real world.

4.4. Evaluation

To evaluate the effectiveness of different classifiers, we selected the messages related to an earthquake captured by SWIM to be the test dataset. Leshan earthquake occurred in January 14, 2015 and SWIM has captured about 300 positive messages about the event. We also selected any messages containing the word "earthquake"around January 14. There were 1,500 messages including the earthquake keyword from 2015-01-13 11:22:03 to 2015-01-15 15:52:44. By examining this test data, we found the positive messages and negative messages were obviously unbalanced. Once the earthquake event occurred, the positive messages emerged sharply, after that, noisy messages (such as the news, comments, preys and so on) would follow. Therefore, in addition to A, P, R, F1, we used ROC curve to determine which classifier is the best.

Classifier	Best size	Recall	F1	Accuracy	Precision
Random forests	2,227	0.867	0.861	0.895	0.832
KNN	2,562	0.779	0.777	0.777	0.780
SVM	2,562	0.858	0.856	0.867	0.846
Naive Bayes	1,138	0.774	0.747	0.844	0.679

Table 2. Classifiers' best Data Size

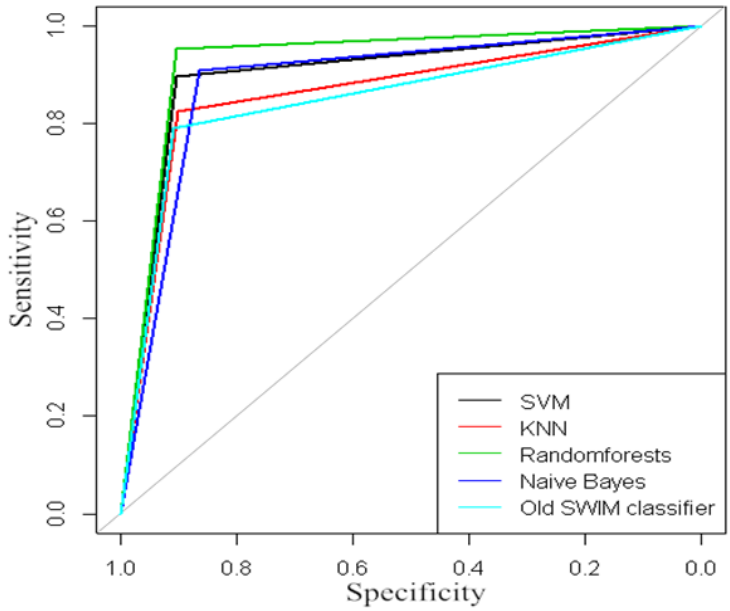

Figure 3. ROC Curves of Four Classifiers

ROC curves show true positive rate on the Y axis, and false positive rate on the X axis. This means that the top and left corner of the plot is the "ideal" point - false positive rate is close to be 0, and true positive rate is close to be 1. It is not realistic to reach the ideal point, but it does mean that bigger area under the curve (AUC), a better classifier.

Figure 3. shows Random forests classifier has the biggest AUC value within the experiment, however this method requires the longest computing time. SWIM catches about 28,000 messages per hour, which means nearly 500 messages per minute. Hence, we performed an experiment in R and calculated the prediction time per 500 messages for each classifier. Since earthquake is an emergency disaster and requires to be reported as soon as possible, from Table 3, we can see that the prediction time of Random forests is a bit hard to satisfy the system's demand. Thus, taking both of performance and computing time into consideration, the new SVM classifier is more suitable for the detection of earthquake messages.

Classifier	A	R	P	F1	AUC	Run Time
SVM	0.902	0.975	0.975	0.937	0.900	15seconds
Random forests	0.912	0.903	0.988	0.944	0.928	88 seconds
KNN	0.886	0.900	0.958	0.928	0.862	8 seconds
Naive Bayes	0.871	0.977	0.863	0.917	0.886	2 seconds
Old SWIM Classifier	0.881	0.898	0.956	0.926	0.850	3 seconds

Table 3. Evaluation results of classifiers

5. Conclusions and Future Work

We developed Sina Weibo Incidents Monitor (SWIM, https://swim.csiro.au/swim/index.html) since September 2014. SWIM monitors Chinese microblogging messages containing disaster keywords and the built-in classifier detects any message which is relating to a disaster just occurred. Since the classifier plays a critical role within SWIM, we always aim to improve its performance. One way to do it is to update its training dataset using the latest data captured by SWIM, and the result shown an incremental improvement to our ability to detect real-event messages. In particular, the F1 score of the new SVM has been improved from 0.926 to 0.937, and more obviously, AUC has been improved from 0.850 to 0.900. In addition to SVM, we also experimented three other algorithms: KNN, Naive Bayes and Random forests. The results show that Random forests has the best Accuracy, Precision, F1 and AUC scores, while it has the longest computing time which makes it less desirable as far as the time is critical.

Future work will include exploring other natural language processing techniques to reduce the computing time of Random forests. Another avenue to explore is to evaluate the performance of classifiers for different types of disasters. In this paper, we only considered earthquake, which needs to be detected as soon as possible, hence

SVM is more suitable than Random forests. However, for less emergency disasters such as heavy rain, other performance measures may be more important.

Furthermore, it is desirable to extract useful information from microblogs relating to a disaster event before, during and after its happening. In particular, we are interested in messages which reporting damage, asking for help, disaster situation, aiding information and public opinion of disasters.

Acknowledgement

The first author of this paper, Hua Bai, was supported by China Scholarship Council to study in Commonwealth Scientific Industrial Research Organisation(CSIRO) Digital Productivity Flagship(DPS) as a visiting PhDÿ thanks for the scholarship.

References

[1] UN OCHA. (2014). World Humanitarian Data and Trends. http://www.unocha.org/data-and-trends-2014/.

[2] Hughes, A.L., Peterson, S., Palen, L. (2015). Social media in emergency management. *FEMA in Higher Education Program*. *In*: Issues in Disaster Science and Management: A Critical Dialogue Between Scientists and Emergency Managers.

[3] Thelwall, M., Stuart, D. (2007). RUOK? Blogging communication technologies during crises. *J. Computer-Mediated Communication* 12 (2) 523-548.

[4] Olteanu, A., Vieweg, S., Castillo. C.(2015). What to expect when the unexpected happens: Social media communications across crises. *In:* Proceedings of Computer Supported Cooperative Work, CSWC, March 2015.

[5] American Red Cross. (2012). More Americans using mobile apps in emergencies. http://www.redcross.org/news/pressrelease/More-Americans-Using-Mobile-Apps-in-Emergencies

[6] Abel, F., Hauff, C., Houben, G. J., Stronkman, R.,Ta K. (2012). Twitcident: Fighting fire with information from social web streams. *In: Proceedings of the 21st International Conference Companion on World Wide Web*, *WWW'12Companion,* p 305-308. ACM.

[7] Power, R., Robinson, B.,Colton, J., Cameron, M. (2014). Emergency situation awareness: Twitter case studies. *In: Proceedings of the 1st International Conference*, p 218-231, Springer International Publishing, October 2014.

[8] Sakaki, T., Okazaki, M., Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering* 25 (4) 919-931.

[9] Robinson, B., Power, R., Cameron, M.(2013). A sensitive Twitter earthquake detector. *In: Proceedings of the 22nd International Conference Companion on World Wide Web, WWW '13 Companion,* p 999–1002, International World Wide Web Conferences Steering Committee.

[10] Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C., Tesconi, M.(2014). EARS (Earthquake Alert and Report System): A real time decision support system for earthquake crisis management. *In:* Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'14, p 1749-1758, ACM, 2014.

[11] Wang, N., She, J., Chen J. (2014). How "Big vs" dominate Chinese microblog: A comparison of verified and unverified users on Sina Weibo. *In:* Proceedings of the 2014 ACM Conference on Web Science, WebSci'14, p 182-186, ACM, 2104.

[12] Gu, Y., Fan, X., Wang, J., Wang, T., Huang, W.(2005). Automatic selection of Chinese stoplist. *Transactions of Beijing Institute of technology* 25 (4) 337-340. (In Chinese).

[13] Li, L., Lee, T., Tang, T., Martin, M. (2013). Mining public opinions from social media in Cantonese. *Overseas Scholars*. Special Volume of 'Big Data' 52-62.

[14] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *In:* Proceedings of the 10th European Conference on Machine Learning, ECML'98, p 137-142, Springer-Verlag.

[15] Burbidge, R., Trotter, M., Buxton, B., Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers and Chemistry* 26 (5) 5-14.

[16] Beyer, K., Goldstein, J., Ramakrishnan, R., et al. When is "*nearest neighbor*" meaningful? *In:* Proceedings of Database Theory—ICDT'99, p 217-235.

[17] Breiman, L. (2001). Random forests. *Machine learning* 45 (1) 5-32.