Iranian J Publ Health, Vol. 43, No.11, Nov 2014, pp.1519-1527

Original Article

Automatic Identification of Messages Related to Adverse Drug Reactions from Online User Reviews using Feature-based Classification

Jingfang LIU [1], Pengzhu ZHANG [1], Yingjie LU [2]

1. *Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China*
2. *School of Economics and Management, Beijing University of Chemical Technology, Beijing, China*

***Corresponding Author:** Email: jingfangliu2014@hotmail.com

(Received 19 July 2014; accepted 04 Oct 2014)

Abstract

Background: User-generated medical messages on Internet contain extensive information related to adverse drug reactions (ADRs) and are known as valuable resources for post-marketing drug surveillance. The aim of this study was to find an effective method to identify messages related to ADRs automatically from online user reviews.

Methods: We conducted experiments on online user reviews using different feature set and different classification technique. Firstly, the messages from three communities, allergy community, schizophrenia community and pain management community, were collected, the 3000 messages were annotated. Secondly, the N-gram-based features set and medical domain-specific features set were generated. Thirdly, three classification techniques, SVM, C4.5 and Naïve Bayes, were used to perform classification tasks separately. Finally, we evaluated the performance of different method using different feature set and different classification technique by comparing the metrics including accuracy and F-measure.

Results: In terms of accuracy, the accuracy of SVM classifier was higher than 0.8, the accuracy of C4.5 classifier or Naïve Bayes classifier was lower than 0.8; meanwhile, the combination feature sets including n-gram-based feature set and domain-specific feature set consistently outperformed single feature set. In terms of F-measure, the highest F-measure is 0.895 which was achieved by using combination feature sets and a SVM classifier. In all, we can get the best classification performance by using combination feature sets and SVM classifier.

Conclusion: By using combination feature sets and SVM classifier, we can get an effective method to identify messages related to ADRs automatically from online user reviews.

Keywords: Adverse drug reaction, Online user review, Feature-based classification

Introduction

"An adverse drug reaction (ADR) is generally defined as an unintended, harmful reaction suspected to be caused by a drug taken under normal conditions" (1). Drug labels contain published ADRs, and ADRs can be found on official websites such as Drugs@FDA (2). However, not all ADRs have been officially identified due to the limitations of controlled clinical trials such as small population size and short duration (3, 4). Therefore, post-marketing drug surveillance has become a very important part of monitoring ADRs (5).

With the development of social media services, some social networking sites such as Medhelp.org

and AskAPatient.com also provide platforms for patients to discuss medications with other patients. These online reviews provided by patients about drugs are widely considered a valuable resource for post-marketing drug surveillance (6). However, these online user reviews are not currently being well utilized by drug regulatory agencies. One important reason is that the online user reviews are full of noise information, highly time-consuming and expensive for the manual identification of the messages related to ADRs out of a large number of unstructured user reviews. So automatic identification of the messages related to ADRs from online user reviews is very challenging but useful research work.

A common problem in social media is noisy data. Filtering noise and extracting relevant data is an important process (7). Owing to the efficient filtering of noise and classification of messages in social media, text classification technologies have been widely used to identify the useful information from huge amounts of unstructured textual data in various domains, including healthcare. Some studies have been performed to extract information related to ADRs from medical text. One study showed that 7.7% of electronic health records included ADRs in their study, and 59% of them could be detected by using their automatic information extraction system (8). Moreover, some research focused on narrative discharge summaries and used the co-occurrence statistics method to extract the relationship between drugs and adverse reactions. However, studies have mainly focused on the context of electronic health records (EHRs) (9, 10), medical case reports (11) and clinical narratives (12-14) that were professionally written medical text. However, the user-generated medical text from social media sites differs significantly from professionally written text. General text classification techniques thus do not produce satisfactory results when identifying message-related ADRs because of the lack of medical knowledge of the users. Because of the rapid development of social media services, researchers have increasingly focused on newly emerged user-generated medical text by patients. Some text classification technologies have been applied to the online user reviews about drugs to extract information related to ADRs. Using controlled medical vocabulary, some studies have used the co-occurrence statistics method (15, 16) or association rules algorithms (17, 18) to identify ADRs from online user reviews and have achieved some progress. However, these statistics-based approaches failed to detect the rare ADRs with low frequencies of occurrence, resulting in a low-precision identification of ADRs (19).

Our aim was to find an effective approach to automatically identify messages related to ADRs from online user reviews using feature-based classification.

Methods

In this study, we used Medhelp.org as our data source. Medhelp.org, one of the most popular online health communities, consists of over 230 discussion boards that concern different disease communities (20). The pain management community is one of the biggest communities. Because the reasons for allergy are diverse, allergy is also among the most frequently discussed topics in online healthcare communities. The cause of schizophrenia is so complicated that the online community becomes a good learning and communication platform for the patient and the kin of the patient. We selected pain management community, the allergy community and the schizophrenia community as data sources. Data collection statistics result is shown in Table 1.

Ethics or Law

In our research, we only use public comments made by users; we do not use any user identification data. Such personal information of the reporting person such as name, age and other demographics are not used or reported as part of the results of the study, therefore, this study does not raise any ethical or legal concern.

We proposed a design framework for the automatic identification of messages related to ADRs from the online user reviews of health communities. We conducted experiments on online user reviews in three communities using different fea-

ture sets and different classification techniques. Our research design consisted of three key steps: message acquisition and annotation, feature set generation, and classification.

The design framework for the automatic identification of messages related to ADRs is shown in Fig. 1. The study flowchart and criteria is shown in Fig. 2.

Message acquisition

In the data collection step, we downloaded the web pages containing user-generated medical messages from three discussion boards. During the setting process of message acquisition, some noisy and unreliable messages were filtered by text preprocessing, including stop words removal and word stemming. We then parsed the pages and stored all user reviews into our database.

Table 1: Data collection statistics

Community name	Messages	Members	Messages per member	Time span
Allergy	9,014	2,203	4.09	September 2008- February 2014
Schizophrenia	1,060	405	2.62	September 2008- February 2014
Pain management	12,180	5,024	2.42	September 2008- February 2014

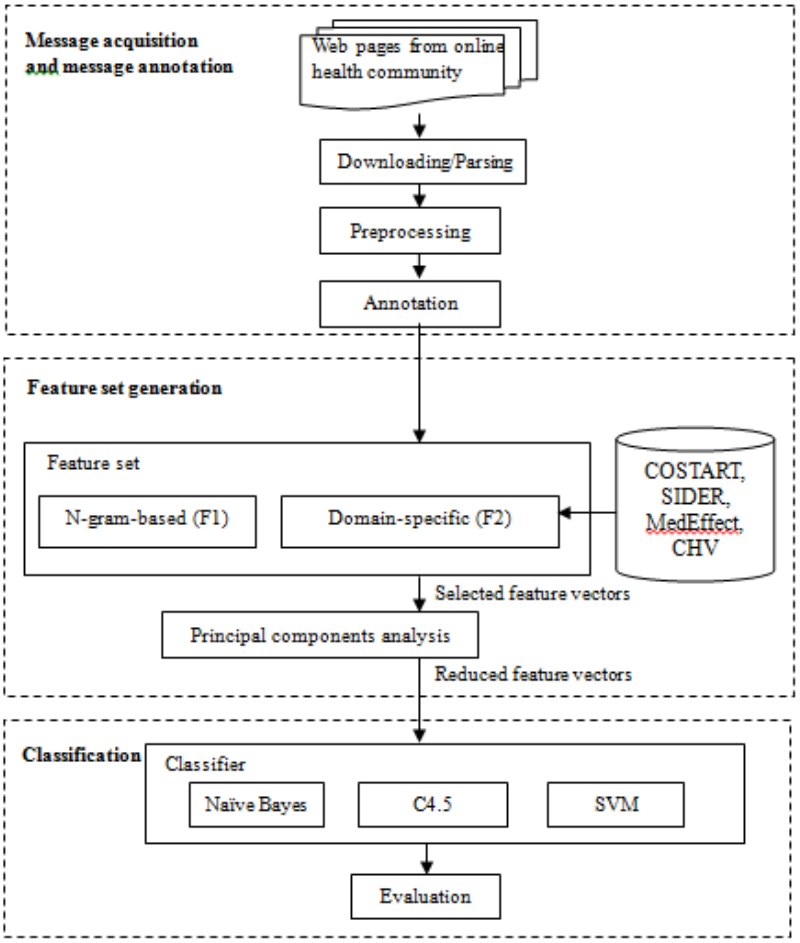

Fig. 1: The design framework for the automatic identification of messages related to ADRs

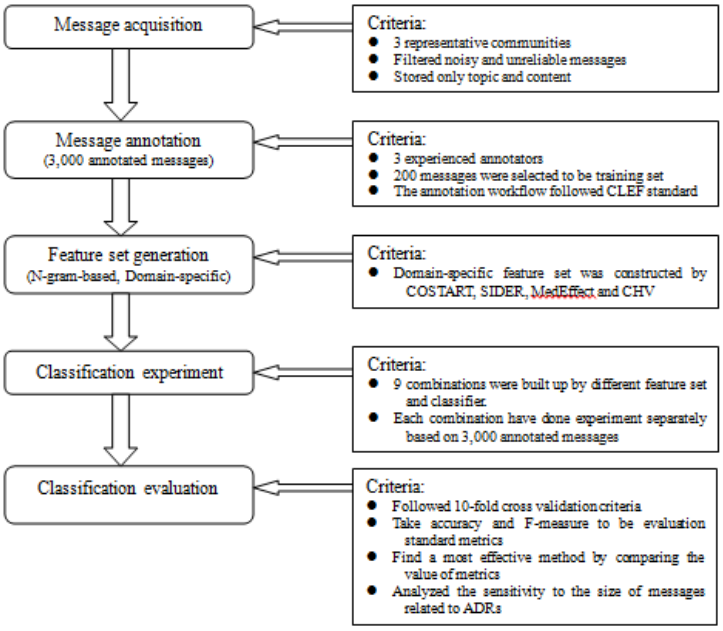

Fig. 2: The study flowchart and criteria

Message annotation

Some messages were independently annotated by three annotators with extensive experience in clinical research and post-marketing drug surveillance. Each message was annotated based on whether ADRs were mentioned or not. A corpus of 3,000 annotated messages is believed to be sufficiently large to support the development and validation of the information extraction system (11). We then randomly selected 3,000 messages from the collected data and manually annotated the messages based on whether ADRs were mentioned or not. The extra 200 messages were selected for use as a training set for annotators. The annotation workflow followed the standards established by the CLEF framework (21).

Feature set generation

Text classification has been widely used to identify messages in biomedical informatics. The vector space model (VSM) was considered an effective method for modeling text content (22). In the VSM, text is represented by a vector of text features. In our study, two types of features were extracted: n-gram-based features and domain-specific features. These features have been adopted in previous text classification studies and proved to be effective for online text classification (23).

Some researchers found that the addition of word n-grams (sequences of words of length n) to text representations indeed improved the performance of text classification (24). So in this study, n-gram-based features were incorporated into the feature set. Studies have found that integrating domain-specific knowledge into textual feature representation could improve classification performance (25). Because user-generated medical messages contain vast medical knowledge, incorporating the medical domain-specific features could significantly enhance the classification performance. We aim to distinguish the messages related to ADRs from online user reviews about drugs, so ADR-related medical terms were introduced into this study as domain-specific features, including COSTART (26), SIDER (27), MedEffect and CHV (28). All synonymous terms are merged into a single unified concept.

To evaluate the effectiveness of our method, we conducted experiments on online user reviews in three communities using different feature sets and different classification techniques. We built three feature sets: feature set F1, feature set F2 and fea-

ture set (F1+F2). F1 features are n-gram-based features; F2 features are domain-specific features; (F1+F2) features are a combination of n-gram-based features and domain-specific features. Feature set F1 was used as the baseline feature set to assess the performance of the other proposed feature sets. The definition of feature set variable is shown in Table 2.

Table 2: Definition of feature set variable

Variable	Value
F1	N-gram-based features
F2	domain-specific features
F1+F2	N-gram-based features and domain-specific features

Classification

In this study, three state-of-the-art classification techniques, SVM, C4.5 and Naïve Bayes, were used to perform classification tasks. To evaluate the effectiveness of our method, we conducted experiments on online user reviews of three communities using different feature sets and different classification techniques. The classifier variable has three values: SVM, C4.5 and Naïve Bayes.

Table 3: Definition of classifier variable

Variable	Value
Classifier	SVM
	C4.5
	Naïve Bayes

Evaluation

Evaluation was performed via 10-fold cross validation. In each fold, 90% messages in the corpus were used as a training set and the remaining 10% messages in the corpus were used as a test set (29). Because the corpus has 3,000 messages, 2,700 messages were considered to be training set and 300 messages were considered to be test set in each fold. We adopted the following standard metrics: accuracy and F-measure to assess the performance of the classification.

$$Accuracy = \frac{number\ of\ all\ correctly\ classified\ messages}{total\ number\ of\ messages} \quad [1]$$

$$F(i) = \frac{2 \times precision(i) \times recall(i)}{precision(i) + recall(i)},\ where\ i = 1,2,\cdots,n \quad [2]$$

Where

$$\text{Precision}(i) = \frac{number\ of\ all\ correctly\ classified\ messages\ for\ class\ i}{total\ number\ of\ messages\ classified\ as\ class\ i}$$

$$\text{Recall}(i) = \frac{number\ of\ all\ correctly\ classified\ messages\ for\ class\ i}{total\ number\ of\ messages\ in\ class\ i}$$

Results

Accuracy evaluation results

We examined the classification accuracy using different feature set and different classifier. From Fig. 3, we could see intuitively that in terms of classifier, the accuracy of SVM classifier was higher than 0.8, the accuracy of C4.5 classifier or Naïve Bayes classifier was lower than 0.8. The SVM classifiers consistently outperformed C4.5 and Naïve Bayes classifiers when using any feature set. Meanwhile, (F1+F2) feature set consistently outperformed other feature set when using any classifier. So in terms of accuracy, the highest accuracy was achieved when using feature set (F1+F2) and SVM classifier.

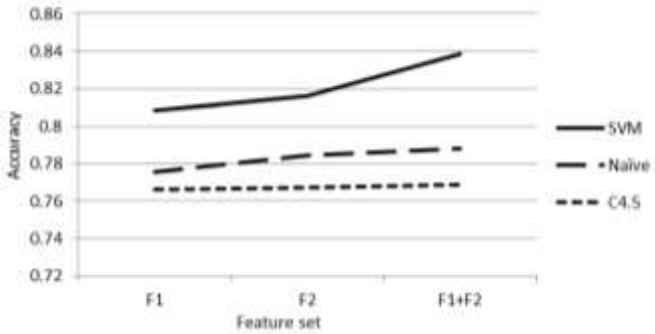

Fig. 3: Accuracy results using different feature sets and classification techniques

F-measure evaluation results

To examine the effects of our classification method, we further evaluated the F-measure of classification results based on different feature set and different classifier. From Table 4, we can see that classification performance using a combination of N-gram-based feature set and domain-specific feature set has significantly outperformed the classification performance using only single feature set, indicating that incorporating feature set enhanced the classification performance significantly. Meanwhile, the F-measure value of SVM classifier was higher than others classifiers. So in term of F-measure, the highest F-measure was achieved when using feature set (F1+F2) and SVM classifier.

Sensitivity analyses

We analyzed the sensitivity of the method to the size of messages related to ADRs. Given a corpus containing x% ADRs messages and (1-x%) non-ADRs messages, we have done experiments when x is 20, 25, 30, 35 and 40.

Table 4: F-measure results using different feature sets and classification techniques

	C4.5	Naïve	SVM
F1	0.595	0.681	0.764
F2	0.559	0.693	0.760
F1+F2	0.648	0.755	0.895

We compared the performance of different feature set on the F-measure value with the size of the ADRs message varies from 20% to 40% of corpus when using SVM classifier. We calculated the p-valure of the paired-sample Wilcoxon signed rank test between the single feature set (F1 or F2) and combination feature sets (F1+F2). The results of sensitivity analyses for feature set are shown in Table 5.

Table 5: The results of sensitivity analyses for feature set

	20% ADR messages (%)	**25% ADR messages (%)**	**30% ADR messages (%)**	**35% ADR messages (%)**	**40% ADR messages (%)**
F1+F2	65.32	70.37	75.66	80.21	89.51
F2	55.49**	61.71**	66.63**	72.97**	80.31*
F1	59.33**	63.99**	69.41**	73.05**	81.72*

P*-values significant at alpha<0.05/*P*-values significant at alpha<0.01

It was found that using (F1+F2) feature set got significantly higher F-measure value than using other feature set. Using (F1+F2) feature set can boost the classification performance than using other feature set, especially when the proportion of ADRs message in corpus is small. Using (F1+F2) feature set can get higher classification performance when the proportion of ADRs message in corpus is large.

We compared performance of different classifier on the F-measure value with the size of the ADRs message varies from 20% to 40% of corpus when using (F1+F2) feature set. We calculated the p-value of the paired-sample Wilcoxon signed rank test between the Naïve Bayes classifier (C4.5 classifier) and SVM classifier. The results of sensitivity analyses for classifier are shown in Table 6.

Table 6: The results of sensitivity analyses for classifier

	20% ADR messages (%)	**25% ADR messages (%)**	**30% ADR messages (%)**	**35% ADR messages (%)**	**40% ADR messages (%)**
SVM	65.32	70.37	75.66	80.21	89.95
Naïve Bayes	61.53**	65.37**	71.59**	76.49**	81.61*
C4.5	59.82**	64.29**	70.18**	75.28**	80.87*

P*-values significant at alpha<0.05/*P*-values significant at alpha<0.01

It was found that the SVM classifier got significantly higher F-measure value than other classifier when using (F1+F2) feature. The SVM classifier can boost the classification performance than other classifier, especially when the proportion of ADRs message in corpus is small. The SVM classifier feature sets can get more effective classification performance when the proportion of ADRs message in corpus is large.

Discussion

It was found that using SVM classifier and combination of N-gram-based feature set and domain-specific feature set is an effective method to identify the messages related to ADRs automatically from online reviews.

Our aim is to distinguish the messages related to ADRs from online user reviews, and the user-generated medical messages contain vast medical knowledge, so it is believed that incorporating the medical domain-specific features could significantly enhance the classification performance. In addition, the combination feature set is a very large feature space in our study. As we known, SVM is commonly believed to best perform in very large feature spaces. The Naïve Bayes and C4.5 can generally have good performance for small or medium feature size. So SVM exhibited better classification performance than other classifiers in our study.

It is inevitable that the potential bias existed in the study process. The three annotators performed annotation process. Because individual knowledge and experience may have some bias, the message annotation may have several misuses. Despite bias, a corpus of 3,000 annotated messages is believed to be sufficiently large to support the development and validation of the information extraction system (11). We choose the messages of three communities as experiment data, the study data can match the research question and the data size is large enough for the study.

The identification results from the three discussion communities using our method is shown in Table 7. The percentage of messages related to ADRs in each community is lower than 32%.

Table 7: Identification result in each community

Community	Total Messages	Messages related to ADRs	Percentage
Allergy	9,014	2,332	25.87
Schizophrenia	1,060	331	31.23
Pain management	12,180	3,349	27.50

After identifying messages related to ADRS, we can save time and speed up efficiency to explore the useful information related to ADRs. Using the method, people can get a large number of messages related to ADRs easily and fast. These messages can provide a lot of valuable ADR reference information for post-marketing drug surveillance. Meanwhile, it is helpful to fully understand ADRs of the drug for patients and drug factories.

The paper also has some limitations that need to be considered further. First, some other features such as sentiment features should be taken into consideration in future studies. Second, there are some other feature reduction methods such as Markov blanket, which was proved effective in some studies. So future researches could explore and compare the performance of different feature reduction methods to obtain the best feature sets. Lastly, other classification techniques, in addition to SVM, C4.5 and Naïve Bayes, should be considered to improve identification performance in future research.

Conclusions

We found that using SVM classifier and combination of N-gram-based feature set and domain-specific feature set is an effective method to identify the messages related to ADRs automatically from online reviews.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

The authors have declared that no competing interests exist.

Acknowledgements

Authors are grateful to the professional help of referee(s) and editorial staff during the process of publishing this study. This research was supported by the National Natural Science Foundation of China under Grant 71301102, China Postdoctoral Science Foundation funded project (2012M5-20903) and Fundamental Research Funds for the Central Universities (ZY1413).

References

1. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G (2010). Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. Proceedings of the 2010 workshop on biomedical natural language processing. Available at *www.google.com*
2. Bellis JR, Kirkham JJ, Nunn AJ, Pirmohamed M (2014). Adverse drug reactions and off-label and unlicensed medicines in children: a prospective cohort study of unplanned admissions to a paediatric hospital. *Br J Clin Pharmaco*, 77(3): 545-553.
3. Munasinghe TJ (2014). Adverse drug reactions: monitoring, reporting and prevention. *Ceylon Med J*, 47(1): 19-21.
4. Aagaard L, Hansen EH (2014). Adverse drug reactions in children reported by European consumers from 2007 to 2011. *Int J Clin Pharmacy*, 36(2): 295-302.
5. Caster O, Conforti A, Viola E, Edwards IR (2014). Methylprednisolone-induced hepatotoxicity: experiences from global adverse drug reaction surveillance. *Eur J Clin Pharmacol*, 70(4): 501-503.
6. Yan P, Chen H, Zeng D (2008). Syndromic surveillance systems. *Annu Rev Inform Sci*, 42(1): 425-495.
7. Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, Holmes JH (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *J Biomed Inform*, 44(6): 989-996.
8. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, Ohe K (2010). Extraction of adverse drug effects from clinical records. *Stud Health Tech Informat*, 160(1): 739-43.
9. Li Y, Salmasian H, Vilar S, Chase H, Friedman C, Wei Y (2013). A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records. *J Am Med Informat Assoc*, 21: 308-314.
10. Wang X, Hripcsak G, Markatou M, Friedman C (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Informat Assoc*, 16(3): 328-337.
11. Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M, Toldo L (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J Biomed Inform*, 45(5): 885-892.
12. Hamon T, Grabar N (2010). Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Informat Assoc*, 17(5): 549-554.
13. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC (2010). MedEx: a medication information extraction system for clinical narratives. *J Am Med Informat Assoc*, 17(1): 19-24.
14. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Informat Assoc*, 18(5): 540-543.
15. Paul A, Jean-Charles D, Gayo D et al. (2013). Design and validation of an automated method to detect known adverse drug reactions in MEDLINE: a contribution from the EU-ADR project. *J Am Med Informat Assoc*, 20(3): 446-52.

16. Jorg H, Dmitry V, Vo Ha N, Shanshan L, Saadat A, Barry L, Robert L, Luis T, Chitta B (2012). A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions. *J Biomed Inform*, 45(5): 842-850.
17. Ji Y, Ying H, Dews P, Mansour A, Tran J, Miller RE, Massanari RM (2011). A potential causal association mining algorithm for screening adverse drug reactions in post-marketing surveillance. *IEEE Trans Inf Technol Biomed*, 15(3): 428-437.
18. Jin H, Chen J, He H, Williams GJ, Kelman C, OKeefe CM (2008). Mining unexpected temporal associations: applications in detecting adverse drug reactions. *IEEE Trans Inf Technol Biomed*, 12(4): 488-500.
19. Benton A, Ungar L, Hill S, Hennessy S, Mao J, Chung A, Holmes JH (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *J Biomed Inform*, 44(6): 989-996.
20. Yang C, Tang X (2010). Who Made the Most Influence in MedHelp? *Intell Syst*, 27: 44–50.
21. Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, Setzer A (2009). Building a semantically annotated corpus of clinical texts. *J Biomed Inform*, 42(5): 950-966.
22. Ning B, Ji Z (2013). Research on web information retrieval based on Vector Space Model. *J NW*, 8(3): 688-695.
23. Alzahrani SM, Salim N, AbrahamA (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Trans Syst Man Cybern C Appl Rev*, 42(2): 133-149.
24. Sidorov G, Velasquez F, Stamatatos E, Gelbukh A, Chanona-Hernandez L (2014). Syntactic n-grams as machine learning features for natural language processing. *Expert Syst Appl*, 41(3): 853-860.
25. Buis LR, Whitten P (2011). Comparison of social support content within online communities for high-and low-survival-rate cancers. *Comput Inform Nu*, 29(8): 461-467.
26. Basch EM, Thaler HT, Shi W, Yakren S, Schrag D (2004). Use of information resources by patients with cancer and their companions. *Cancer*, 100(11): 2476-2483.
27. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010). A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*, 6: 343.
28. Zeng QT, Tse T (2006). Exploring and developing consumer health vocabularies. *J Am Med Informat Assoc*, 13(1): 24-29.
29. Fan J, Guo S, Hao N (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J R Stat Soc Series B Stat Methodol*, 74(1): 37-65.