International Journal of u- and e- Service, Science and Technology
Vol.8, No.3 (2015), pp.239-246
http://dx.doi.org/10.14257/ijunesst.2015.8.3.23

A Hybrid Public Opinion Analysis Method Based on Improved Clustering and Mutual Information

Zhiqiang Geng*, Xia Tang, Yikang Zhang and Yongming Han

(College of information science and technology, Beijing University of chemical technology, Beijing, 10029, China)
gengzhiqiang@mail.buct.edu.cn,2012210217@grad.buct.edu.cn,
2013200807@grad.buct.edu.cn, hanym@mail.buct.edu.cn

Abstract

The Internet is frequently used as a medium for exchange of information and opinions, and it is imperative to conduct public opinion analysis to get people's opinions well understood and guided. In this paper a hybrid public opinion analysis method based on improved clustering and mutual information is proposed. During feature extraction, the weights of words are modified based on Part-of-Speech Tagging to reduce the dimensions of original texts. As for clustering, a novel density peak algorithm is improved and combined with binary search algorithm to determine the cluster number K and initial centers for KMeans. Then hot words extraction, sentiment analysis and trend analysis for each cluster are processed with mutual information to mine useful knowledge to help decision-making. Extensive experiments are conducted on Hadoop, and the results show that our hybrid Public Opinion Analysis method is quite effective and has certain significance.

Keywords*: clustering, mutual information, trend analysis*

1. Introduction

The Internet is becoming a spreading platform for public opinion partly induced by the emergence of social networks, blogs, wiki and other online media. The Public Opinion analysis focuses on analyzing people's opinions, sentiment, and attitudes towards entities such as organizations, individuals and events, which involve natural language processing, data mining, sentiment analysis, etc. It is important to grasp public opinion on the Internet over time and understand its trends to help users make decisions quickly and correctly. Public Opinion analysis includes data collection, data preprocessing, topic clustering, hot words extraction, sentiment analysis and trend analysis. The current Public Opinion analysis with massive news is too difficult to be processed by traditional methods and platforms. The MapReduce[1] is a popular framework for processing large datasets in parallel over a cluster. It has gained wide attention for its high scalability, reliability and low cost, simplicity and flexibility. Hadoop[2], a popular open source implementation of MapReduce, has been successfully used in various applications such as data mining and web processing.

In the field of Public Opinion analysis, predecessors have done a lot of work and accumulated rich experience. Gao constructs a three-dimensional space combined with information space model based on the mechanism of opinion information dissemination. The Public Opinion detection index system is set up by analytic hierarchy process [3], which offers quantitative calculation method of public opinion index but lacks of a large number of statistics to test and verify. ICTCLAS, a Chinese word segmentation method, has been widely used for its high accuracy and part-of-speech tagging. Moghaddam proposes an improved method which offers a filter based on pattern to remove useless words [6]. Algorithms based on word networks, such as SWN model algorithm, BC

ISSN: 2005-4246 IJUNESST

algorithm [7], can extract high frequency words smoothly, but can't take out words with important or low frequency on document level. On clustering methods, Sajib proposes a clustering method called "mine the easy, classify the hard" [8], which can automatically mark comments to obtain effective opinions and use these results to classify the fuzzy comments. While the clustering process is not just designed for sentiment classification, it can be used for other text classification tasks. In terms of sentiment analysis, Qiu[9] puts forward a method of two-way transmission which can get specific sentiment dictionary and direction set in specific areas and extract an amount of new sentiment words.

In this paper, we propose a hybrid Public Opinion analysis method based on improved clustering and mutual information. We mainly make following contributions:

- In order to solve the problems of KMeans algorithm, we improve the density peak algorithm proposed in [16] and apply binary search algorithm to decide the dc value, which further decide the cluster number K and initial centers for KMeans.
- In terms of feature selection, useless words are filtered by Part-of-Speech tagging and stop-dictionary. The weights of candidate words are computed according to Part-of-Speech.
- Mutual information is used to extract hot words and calculate positive\negative sentiment score for each cluster, and its overall sentiment trend is analyzed to help users make decisions quickly and correctly.

2. Text Clustering of News

2.1. Feature Extraction of News

A web crawler is used in this paper, based on breadth-first traversal, to obtain news links. We use the htmlparser to extract relevant stories, which include titles, contents, authors and published time. The contents and titles of considered text are used for clustering.

The most difficult problem of feature selection is the "dimension disaster", where text content is mapped to a high-dimensional vector space after word segmentation. How to reduce the dimensions effectively is a key problem, otherwise it would put a bad impact on the final result of clustering. Therefore, before features are extracted, ICTCLAS is used to conduct the Chinese word segmentation. Then stop-dictionary and Part-of-Speech tagging are applied to filter useless words. A stop-dictionary contains many words widely used but not significant for clustering. By filtering the words contained in stop-dictionary, the original word collection can be greatly reduced. Lastly, the words with their POS (Part Of Speech) being noun, adjective or verb are kept, as they can make topics stand out better.

Finally, the words in each news document are saved in the VSM (Vector Space Model)[10]. The TF-IDF is widely used to compute the weight of each word in the VSM. In a news document, terms can show their importance by its frequency. For a word t_i, its importance can be calculated as formula 1.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$$idf_i = -qlogq - (1-q)log(1-q), \text{ where } q = \frac{D_{t_i}}{|D|} + 0.01 \tag{2}$$

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

where $n_{i,j}$ represents the occurrence of the word t_i in the text d_j, the denominator is the entire occurrence of all words in d_j. Usually, the Inversed Text Frequency (IDF) is computed by the number of all documents dividing the number of documents containing word t_i. As the words whose document frequency are too high or low cannot contribute

much for clustering, therefore, an improved IDF formula proposed in an existing paper[18] is adopted as formula 2, where D represents the number of all documents, and D_{ti} represents the number of document containing word t_i. The TF-IDF value is calculated as formula 3.

After above phases, each news document can be represented as a TF-IDF vector, with each dimension represents the importance of the corresponding word. As TF-IDF only indicates statistical information of a single word and does not contain its inherent attributes and position in a document, it is found that the clustering effect is not good enough using TF-IDF directly. Therefor, before clustering, we adjust TF-IDF as follows.

If a word belongs to the specific noun, its weight changes to 5 times of its original value;

If a word belongs to the common noun, its weight changes to 2 times of its original value;

If a word appears in the title, its weight changes to 2 times of its original value.

2.2. News Clustering

The KMeans clustering algorithm[11] is not only simple but can be implemented in parallel. However, it is difficult to determine the proper clustering number K and choose appropriate initial centers. In this paper, a novel method to decide K and initial centers is referred to[16], its main idea is as follows. Assume the cluster centers are surrounded by neighbors with lower local density and they are at a relatively large distance from any points with a higher local density. Each sample is characterized by two attributes: local density ρ_i and its distance δ_i from the points of higher density, which are obtained as formula 4 and formula 5, where *dc* is a cut-off distance. Both of the quantities depend only on the distances d_{ij} between two points, assuming to satisfy the triangular inequality. By observing the decision graph with ρ_i *as its x axis and* δ_i as its y axis, the data points whose ρ_i and δ_i are both large are chosen as the initial centers and the number of these points is the final K.

$$\rho_i = \sum_i \chi(d_{ij} - d_c) \quad \text{, where} \quad \chi(x) = \begin{cases} 1 & if x < 0 \\ 0 & otherwise \end{cases} \tag{4}$$

$$\delta_i = \min_j : \rho_j > \rho_i(d_{ij}) \tag{5}$$

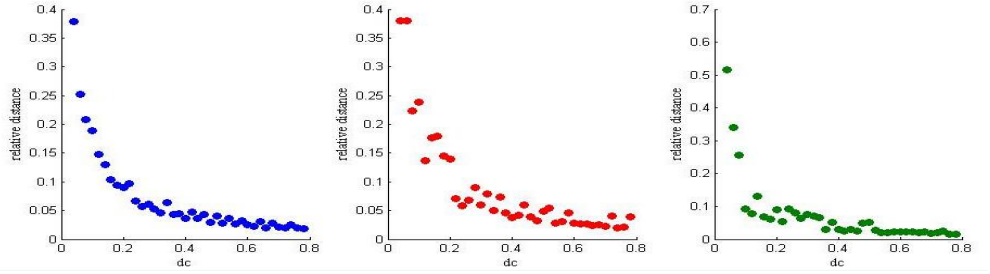
However, in experiments, we find the initial centers decision graph is quite different with different *dc*, and no accurate algorithm to determine *dc* is presented in [16]. An improving algorithm is proposed for this problem, its main idea is as follows. We use binary search algorithm to find the best dc. Given current value range of $dc[c_{left}, c_{right}]$, we compute its relative distance rd_{left}, rd_{right}, rd_{mid} according to formula 6, when dc equals to c_{left}, c_{right} and $c_{mid}=(c_{left}+c_{right})/2$ respectively. If the difference between rd_{mid} and rd_{right} is quite larger, the real *dc* is between c_{mid} and c_{right}, otherwise it is between c_{left} and c_{mid}. We continue this search process until the left difference equals to the right difference. To get better understood, in Figure1 we draw the points with *dc* as its *x* axis *and* relative distance as its *y* axis. Actually, the best *dc* is around the inflection point.

$$rd(dc_i) = \frac{1}{|\overrightarrow{\rho_{dc=c_i}}|} \tag{6}$$

,where $\overrightarrow{\rho_{dc=c_i}}$ represents the vector composed of all ρ_i when *dc* equals to c_i.

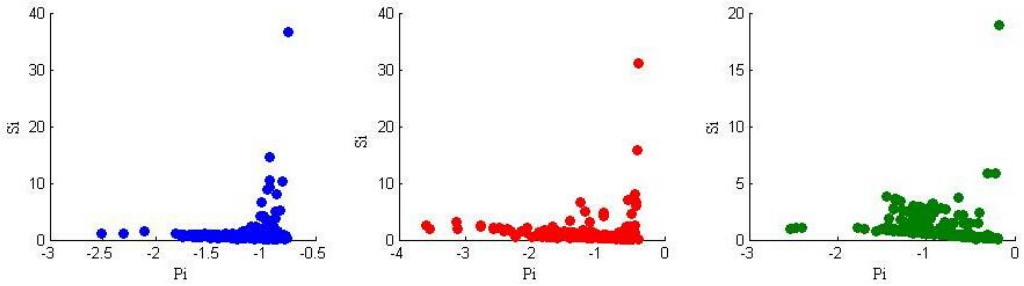
After we get the final *dc*, the initial centers decision graph with it for dataset I, II and III are shown in Figure 2, where the data points whose ρ_i and δ_i are both large are chosen as the initial centers and the number of these points as the final K. When applied to the massive datasets, it is improper to directly draw all the points in the initial center decision graph, so

only the top 20% points with largest Euclidean distance to the origin are drawn.



1(a) dataset I 1(b) dataset II 1(c) dataset III

Figure 1. The Relation of Different *dc* and its Relative Distance for Different Dataset



2(a) dataset I 2(b) dataset II 2(c) dataset III

Figure 2. The Initial Centers Decision Graph for Different Datasets

2.3. Hot Words Extraction

The hot words can indicate the major content of each cluster, so mutual information of a word and the cluster where it appears are calculated in the process of extracting hot words, as is shown in formula 7.

$$I(a_i;b_j)=\log\frac{p(a_i\mid b_j)}{p(a_i)}=\log\frac{p(a_i,b_j)}{p(a_i)p(b_j)} \tag{7}$$

,where a_i represents a word, b_j represents a clustering, $p(a_i, b_j)$ is the number of documents which contains a_i in cluster b_j. After computing mutual information for all words in the cluster b_j, the top T words with largest mutual information are chosen as hot words. Through extensive experiments, we find it is proper when T is 5.

3. Sentiment Analysis

3.1 Sentiment Score

The sentiment analysis of news mainly focuses on discovering what opinion is held by people, the level of the sentiment, and the trend of their sentiment over time. In this paper, a dictionary from WordNet [13] is used to help sentiment analysis. Each word's polar score (positive or negative) is calculated directly by their TF value as we think if a positive(or negative) word appears many times it can measure the level of sentiment score greater. By summing the polar score of all positive and negative words respectively, the overall opinion of a certain topic or cluster (positive or negative) can be obtained. Specifically, positive and negative score of a cluster are calculated as the following formula 8 ~11.

$$\text{Pos\_score(filei)} = \left[\frac{\sum_{i=1}^{pos} Intensity(i)}{pos} \right] \tag{8}$$

$$\text{Neg\_score(filei)} = \left[\frac{\sum_{i=1}^{neg} Intensity(i)}{neg} \right] \tag{9}$$

$$Pos\_score(clusterj) = \sum Pos\_score(i) \tag{10}$$

$$\text{Neg\_}score(clusterj) = \sum Neg\_score(i), i \in clusterj \tag{11}$$

3.2. Trend Analysis

The trend analysis is used to describe people's opinions of a topic over time. First of all, other information, such as author, source, publish time should be combined with the content of a document, and sorted by publish time to clearly help user make decisions. In addition, the level of a kind of attention is computed by summing the sentiment score of the same topic in a variety of sources over time.

4. Experiments and Evaluation

4.1. Datasets

20,000 political news are obtained from *china.com*[14], which include many different topics, including China-US relation, China-Japan relation and relation between China and Russia, etc. They may also occur in the same news content, which increases the difficulty during clustering.

4.2. Environment

Extensive experiments are conducted over Hadoop to evaluate the approaches proposed in this paper. Our cluster contains 1 master node and 9 slave nodes. Each node is a 64-bit Intel Core machine with one four-core CPU 2.0 Ghz CPU, 4GB physical memory and 250GB disk, and runs CentOS release 6.4. The version of Hadoop is Apache Hadoop 1.2.0. Each node is configured with 4 map slots and 2 reduce slots. Preprocessing includes merging all documents to a large file, Chinese word segmentation and Part-of-Speech tagging. Then Feature extraction, clustering, hot words extraction and sentiment analysis, trend analysis are conducted. The final results are saved in the HDFS file system whose output format is easy to understand. Hadoop is suitable for processing large files [15], many small files will greatly consume memory and lead to excessive map tasks. In this paper a job is submited for merging small files into a single sequence file. The key of each record is the title of a news document, and the value is the text content of the news document in form of word segmentation.

4.3. The Results of Experiments

4.3.1. Clustering: After clustering, each cluster may still contain more than one topic, which means these clusters are on the verge of several clusters, which are related to multiple clusters. Therefore, when evaluating the clustering results, the topic with the largest ratio is chosen as the theme of a cluster. Table 1 shows the clustering result of all news documents, the average accuracy rate and recall rate reach 88.70% and 83.60%

respectively. Our experiment results show that the appropriate clustering number is 32. Specifically, Table 2 is part of clustering results about the "diaoyu islands" issue, the accuracy and recall rate of this topic are as high as 99.3% and 98.2%.

4.3.2. The Results of Hot Words Extraction: The result of hot words extraction for all news is shown in Table 3, it can be seen that hot words do express quite much information. For most of clusters, hot words can reflect the real topic of a cluster. However, the results of hot words extraction for a few clusters are not good enough. In summary, 28 out of 32 clusters have satisfying results.

Table 1. Clustering Results of 20000 Documents

Iterations	Documents	Accuracy Rate	Recall Rate
50	20000	88.70%	83.60%

Table 2. Parts Clustering Results of the "Diaoyu Islands" Issue

Cluster id	Title	Published time	Source	Positive senstiment score	Negative sentiment score
17	新驻华大使是日本财界的大人物.txt	2010-7-26	中国网	s	4.284
17	一意孤行，日本将付出沉重代价.txt	2010-9-20	中国网	1.984	0.249
17	促进中日友好，综合发力比盯首相面孔更重要.txt	2011-9-1	中国网	5.481	0.015
17	日本欲拿南海先练手.txt	2011-9-13	中国网	1.459	0.326
17	日本军演明确中国为假想敌.txt	2011-11-11	环球时报	1.341	0.692
17	日本因何推崇中国留学生的漫画？.txt	2012-2-10	中新网	0.857	0.046
17	否定南京大屠杀凸现日本政坛之怪现象.txt	2012-2-22	民网-国际频	1.145	0.42
17	钟之乌礁变岛，日本欲利用“国际规则”建岛圈海.txt	2012-5-29	中国网	0.924	0.408
17	“声外击内”能挽救野田政权危局？.txt	2012-7-13	中国网	2.608	0.056
17	一种情绪多种应对.txt	2012-9-3	京华时报	2.411	0.847
17	“操日本”串红，爱国不应有“看客”心态.txt	2012-9-12	中国网	0.229	2.784
17	候任日本驻华大使之死的三重警示.txt	2012-9-17	中国网	0.604	2.488
17	行动表达不应失去底线.txt	2012-9-17	长江日报	2.752	0.021
17	东北亚对日本的反包围圈已悄然形成.txt	2012-9-18	中国网	0.356	0.274
17	只要日本死不悔改，经济制裁就不能停止.txt	2012-9-20	千龙网	9.294	0.163
17	只要日本政府死不悔改，制裁就不能停止.txt	2012-9-20	千龙网	5.8567	0.163
17	日本政府不悔改，制裁就不能停止.txt	2012-9-20	中国网	5.654	0.157
17	日本巨额金援撬动巴拿马在钓鱼岛站队.txt	2012-10-25	中国网	3.199	0.288
17	日本地方议会的喊叫衰渎了中韩.txt	2012-10-30	中国网	2.0446	0.139

Table 3. Parts of Hot Word Extraction Results

Cluster id	Hot words
4	(爱国/a 理性/n 日货/n 简单/a 配合/v)
8	(菲律宾/nsf 中国/ns 紧张/a 精力/n 值得/v)
20	(白皮书/n 结局/n 佳彦/nr2 个体/n 坚韧/an)
19	(阿里/ns 表演/v 调停/vi 节奏/n 求和/vi)
15	(中国海/ns 受挫/vi 牢记/v 盘点/n 训练/vn)
6	(克尔/nsf 默克尔/nrf 习近平/nr 样板/n 眼光/n)
0	(美国/nsf 华为/nz 中兴/nz 国会/n 党委/n)
2	(保护主义/n 澳大利亚/nsf 包容/vn 包容/v 蔓延/vi)
17	(巴拿马/nsf 漫画/n 南京大屠杀/nz 泡沫/n 召回/v)
13	(湄公河/nsf 中国/ns 公民/n 大局/n 海外/s)
11	(韩国/nsf 中国/ns 海警/n 韩方/nr 人权/n)
14	(巡视/v 私有化/v 私有化/vn 普京/nrf 公关/n)
7	(越菲/nz 索马里/nsf 普京/nrf 太平/a 油气/n)

4.3.3. The Results of Sentiment Analysis: Statistics results for overall sentiment level of all clusters are shown in Figure 3. All clusters are more inclined to express positive feelings. Figure 4 shows detailed sentiment trend of the "diaoyu islands" issue, we can see the positive sentiment score is higher than the negative sentiment score over time. However, on days near September 18, 2012, the negative sentiment score increased quickly and the positive sentiment fell sharply, which may be related to 9.18 issue, so some proper and effective measures should be taken to prevent potential problems.

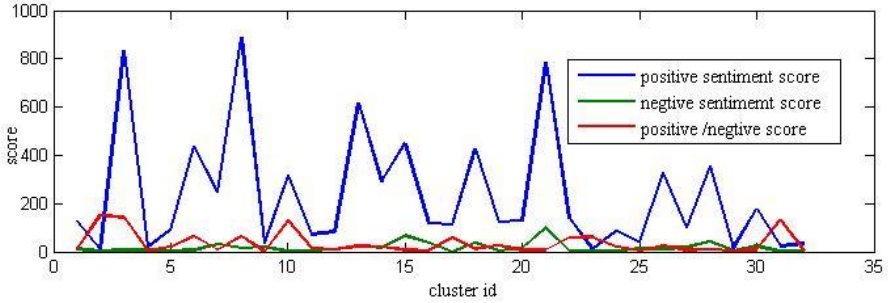


Figure 3. The Positive and Negative Sentiment Score of Clusterings

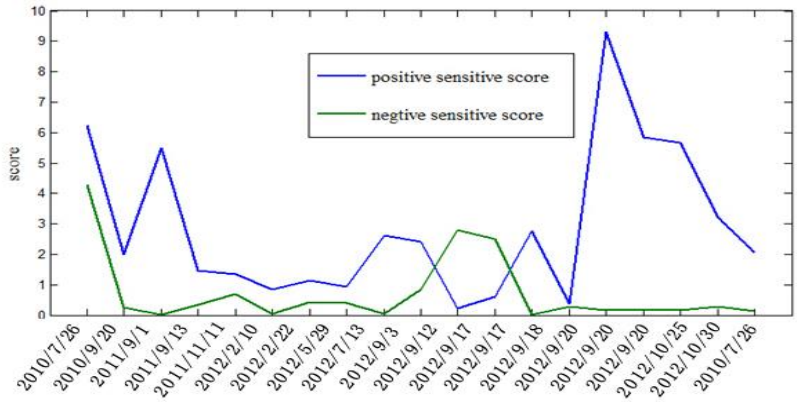


Figure 4. Positive and Negative Sentiment Trend of the "Diaoyu Islands" Issue

5. Conclusion and Future Work

In this paper a hybrid Public Opinion analysis method is proposed to solve the problems of feature selection and deciding the cluster number K and initial centers for KMeans. Stop-dictionary, the Part-of-Speech tagging and setting different weights of candidate words according to different parts of speech are used to select useful features. The initial centers decision graph and binary search algorithm are applied to decide the cluster number K and initial centers. The mutual information is computed for each cluster to extract hot words, and the positive and negative sentiment score of each cluster is calculated in sentiment analysis phase. Finally, the sentiment change trend for each cluster is analyzed to discover people's opinions tendency and help decision-making. In the future work, in order to conduct more comprehensive analysis, the amount of collected news will be expanded. Only relying on mutual information to extract hot words can lead to inaccurate results sometimes, so a variety of related algorithms should be studied to improve the hot words extraction results.

Acknowledgement

This research was partly funded by National Natural Science Foundation of China (61374166), the Doctoral Fund of Ministry of Education of China (20120010110010) and the Fundamental Research Funds for the Central Universities (YS1404).

References

[1] J. Dean and S. Ghemawat, COMMUN. ACM., vol. 1, no. 51, **(2008)**.
[2] Hadoop, Apache Software Foundation, http://hadoop.apache.org.
[3] C. Gao, X. Rong and Y. Chen, Journal of information, vol. 9, no.30, **(2011)**.

[4] A. Younus, M. A. Qureshi, F. F. Asar,"What do the average twitterers say: A twitter model for public opinion analysis in the face of major political events", Advances in Social Networks Analysis and Mining (ASONAM), **(2011)**; Kaohsiung, Taiwan.
[5] H. P. Zhang, H. K.Yu, D. Y. Xiong, HHMM-based Chinese lexical analyzer ICTCLAS. Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics, **(2003)**; Stroudsburg, PA, USA.
[6] S. Moghaddam, M. Ester, "Opinion digger: an unsupervised opinion miner from unstructured product reviews", Proceedings of the 19th ACM international conference on Information and knowledge management, **(2010)**; Toronto , ON, Canada.
[7] L. Ma, L. Jiao, B. Lin and Y. Xhou, "Journal of Chinese Information Processing", vol. 5, no. 121, **(2009)**.
[8] S. Dasgupta, V. Ng, "Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification", Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, **(2009)**; Suntec, Singapore.
[9] G. Qiu, B. Liu, J. Bu, "Expanding Domain Sentiment Lexicon through Double Propagation", IJCAI, **(2009)**; Pasadena, California.
[10] D. L. Lee, H. Chuang, K. Seamons, Software, vol. 2, no. 14, **(1997)**.
[11] A. K. Jain, Pattern Recognition Letters, vol. 8, no. 31, **(2010)**.
[12] A. Rajaraman, J.D. Ullman, "Mining of massive datasets", Cambridge University Press, **(2011)**.
[13] G. A. Miller, Communications of the ACM, vol. 11, no. 38, **(1995)**.
[14] http://www.china.com.cn/
[15] J. Shafer, S. Rixner and A. L. Cox, "The Hadoop distributed filesystem: Balancing portability and performance. Performance Analysis of Systems & Software (ISPASS)", **(2010)**; White Plains, NY.
[16] A. Rodriguez, A. Laio, Science, vol. 6191, no. 344, **(2014)**.
[17] http://cs.joensuu.fi/sipu/datasets/
[18] D.D Xu , S.B.Wu, "An Improved TFIDF Algorithm in Text Classification", Applied Mechanics and Materials**, (2014)**; Changsha, China.

