Advances in Environmental Biology, 11(5) May 2017, Pages: 19-25

AENSI Journals

Advances in Environmental Biology

ISSN-1995-0756 EISSN-1998-1066

Journal home page: http://www.aensiweb.com/AEB/

A Novel Sentiment Analysis Technique in Disease Classification

[1]K. Anita Davamani, [2]Dr. C. R. Rene Robin, [3]Kamatchi. S, [4]Krithika. S. R, [5]Manisha. P, [6]Santhosh. T

[1]Asst. Professor, Dept of CSE, Jerusalem College of Engineering, Chennai – 100.
[2]Professor & Joint Director, Dept of CSE, Jerusalem College of Engineering, Chennai – 100.
[3-6]UG Students, Dept of CSE, Jerusalem College of Engineering, Chennai – 100.

Address For Correspondence:
K. Anita Davamani, Asst. Professor, Dept of CSE, Jerusalem College of Engineering, Chennai – 100.
E-mail: anitadavamani@gmail.com

Received 8 January 2017; Accepted 28 April 2017; Available online 24 May 2017

ABSTRACT
Online Health Communities (OHCs) provide useful source of information and support for health seekers. The Challenge organized by the Informatics for Integrating Biology and the Bedside (I2B2), a National Center for Biomedical Computing, asked participants to construct systems that could correctly replicate the textual and intuitive judgments of the medical experts on disease category based on narrative patient records. Since hospitals usually store a considerable amount of information (patient data) as free text, such systems have a great potential in aiding research on disease due to their capability to process large document repositories both cost and time efficiently. In our work, a dictionary consisting of unique words collected from health care centers is proposed and these words are trained using Naïve Bayes Algorithm based on the Probability Theory. To classify the disease based on the symptoms posted by the health seekers or patients, we have extracted each word of the posting using Tokenization method and then match those words with dictionary words for classification. Finally testing performed against 500 posts and shows promising result.

KEYWORDS: Online Health Communities(OHCs), Sentiment Analysis, Naïve Bayes, Tokenization, Probability Theory, Disease Classification.

INTRODUCTION

Sentiment Analysis, also known as opinion mining, involves building a system to collect and categorize opinions about a product. Automated opinion mining often uses machine learning, a type of artificial intelligence (AI), to mine text for sentiment. Opinion Mining and Sentiment Analysis is a Natural Language Processing (NLP) technique that automatically extracts the opinion, sentiments, attitude, emotions, views etc., in proper context and classify these into different categories like positive, negative , neutral etc. Other terms used for this research domain are subjectivity analysis, subjectivity detection, appraisal extraction and review mining, sentiment mining.

The two important tasks involved in Opinion Mining and Sentiment Analysis are [1] Opinion Extraction: extracting the opinionated phrases, in proper context, from free text and [2] Sentiment classification: classifying opinionated phrases based on sentiment orientation.

Nowadays, accelerated urbanization and improving living standard have brought some unexpected negative influences, making modern citizens more suffered from chronic diseases. On the other hand, with the innovation of information technology, Web 2.0 is able to provide an accessible and unobstructed channel of communication, which attracts more patients to seek support of health problems through the Internet. Therefore, increasing researches have been focusing on improvement of effective management of online health community. It should

been mentioned that community activities reflect the relationship and communication of people, which involves more or less emotional elements. Further, this may influence the management of the entire community.

The Mission of e health-care is to help patients, physicians, and community hospitals to make appropriate use of information and communication technologies in order to improve access and quality of health care delivery and reduce the cost of its management. E health-care Foundation supports civil, community, government, and non-government health care delivery vehicles in improving their management by making available web-based patient care, physician, and hospital management solutions affordably. It supports exchange of patient data in a secure network across all the stakeholders in delivering health care.

Sentiment analysis for health care deals with the diagnosis of health care related problems identified by the patients themselves. It takes the patients opinions into perspective to make policies and modifications that could directly address their problems. Sentiment analysis is used with commercial products to great effect and has outgrown to other application areas. Aspect based analysis of health care, not only recommend the services and treatments but also present their strong features for which they are preferred. Machine learning techniques are used to analyze millions of review documents and conclude them towards an efficient and accurate decision. The supervised techniques have high accuracy but are not extendable to unknown domains while unsupervised techniques have low accuracy. More work is targeted to improve the accuracy of the unsupervised techniques as they are more practical in this time of information flooding.

Sentiment analysis in health care is an emerging trend that can give health care organizations a competitive edge in understanding and improving the patient experience. Sentiment analysis in health care uses natural language software to categorize and assess written and spoken comments by patients about their health care experience. When combined with patient satisfaction data, sentiment analysis provides health care organizations with much deeper insight into patient perceptions and with an understanding of where changes can have the most dramatic impact on improving patient experience.

II. Sentiment Analysis Techniques:

A. Machine Learning:

Machine learning based Sentiment Analysis or classification can be done in two ways: 1) Sentiment Analysis by using supervised machine learning techniques and 2) Sentiment Analysis by using unsupervised machine learning techniques.

1) Supervised Machine Learning:

In Supervised Machine learning techniques, two types of data sets are required: training data set and test data set. An automatic classifier learns the classification factors of the document from the training set and the accuracy in classification can be evaluated using the test set. Various machine learning algorithms are available that can be used very well to classify the documents. The machine learning algorithms like Support Vector Machine (SVM), Naive Bayes (NB) and maximum entropy (ME) are used successfully in many research and they performed well in the sentiment classification.

2) Unsupervised Machine Learning:

Lexicon Based Method is an Unsupervised Learning approach since it does not require prior training data sets. It is a semantic orientation approach to opinion mining in which sentiment polarity of features present in the given document are determined by comparing these features with semantic lexicons. Semantic lexicon contains lists of words whose sentiment orientation is determined already. It classifies the document by aggregating the sentiment orientation of all opinion words present in the document, documents with more positive word lexicons is classified as positive document and the documents with more negative word lexicons is classified as negative document.

B. Hybrid Technique:

Some researchers combined the supervised machine learning and lexicon based approaches together to improve sentiment classification performance. They considered both general purpose lexicon and domain specific lexicon for determining polarity orientation of sentiment words and feed these lexicons into supervised learning algorithm, SVM. They found that general purpose lexicon performed very poor while domain specific lexicon performed very well. The system classified the sentiment in two steps: First the classifier is trained to predict the aspects and In Next the classifier is trained to predict the sentiments related to the aspects collected in step1.Their system yielded around 66.8% accuracy.

III. Related Works:

Sentiment mining aims at extracting features on which users express their opinions in order to determine the user's sentiment towards the query object. We mine over 70 million Twitter micro-blogs to gain knowledge regarding tourist sentiment on the travel resort destination Cancun in the Yucatan Peninsula of Mexico. We measure sentiment using a binary choice keyword algorithm and a multi-knowledge based approach is proposed

using, Self- Organizing Maps and tourism domain knowledge in order to model sentiment [3]. We develop a visual model to express this taxonomy of sentiment vocabulary and then apply this model to maximums and minimums in the time sentiment data. The results show practical knowledge can be extracted.

Nowadays, classifying sentiment from social media has been a strategic thing since people can express their feeling about something in an easy way and short text. Mining opinion from social media has become important because people are usually honest with their feeling on something. In our research, we tried to identify the problems of classifying sentiment from Indonesian social media[4]. We identified that people tend to express their opinion in text while the emoticon is rarely used and sometimes misleading. We also identified that the Indonesian social media opinion can be classified not only to positive, negative, neutral and question but also to a special mix case between negative and question type. Basically there are two levels of problem: word level and sentence level. Word level problems include the usage of punctuation mark, the number usage to replace letter, misspelled word and the usage of nonstandard abbreviation. In sentence level, the problem is related with the sentiment type such as mentioned before. In our research, we built a sentiment classification system which includes several steps such as text pre-processing, feature extraction, and classification. The text pre-processing aims to transform the informal text into formal text. The word formalization method in that we use is the deletion of punctuation mark, the tokenization, conversion of number to letter, the reduction of repetition letter, and using corpus with Levensthein to formalize abbreviation. The sentence formalization method that we use is negation handling, sentiment relative, and affixes handling. Rule-based, SVM and Maximum Entropy are used as the classification algorithms with features of count of positive, negative, and question word in sentence and bigram. From our experimental result, the best classification method is SVM that yields 83.5% accuracy.

A typical method to obtain valuable information is to extract the sentiment or opinion from a message. Machine learning technologies are widely used in sentiment classification because of their ability to "learn" from the training dataset to predict or support decision making with relatively high accuracy. However, when the dataset is large, some algorithms might not scale up well. The scalability of Naive Bayes classifier (NBC) in large datasets is evaluated [5]. Instead of using a standard library (e.g., Mahout), we implemented NBC to achieve fine-grain control of the analysis procedure. A Big Data analysing system is also design for this study. The result is encouraging in that the accuracy of NBC is improved and approaches 82% when the dataset size increases. We have demonstrated that NBC is able to scale up to analyze the sentiment of millions movie reviews with increasing throughput.

SNS is one of the most effective communication tools and it has brought about drastic changes in our lives. Recently, however, a phenomenon called flaming or backlash becomes an imminent problem to private companies. A flaming incident is usually triggered by thoughtless comments/actions on SNS, and it sometimes ends up damaging to the company's reputation seriously. In this paper, in order to prevent such unexpected damage to the company's reputation, a new approach is proposed to sentiment analysis using a Naıve Bayes classifier, in which the features of tweets/comments are selected based on entropy-based criteria and an empirical rule to capture negative expressions [6]. In addition, we propose a semi-supervised learning approach to relabeling noisy training data, which come from various SNS media such as Twitter, Facebook, blogs and a Japanese textboard called '2-channel'. In the experiments, we use four data sets of users' comments, which were posted to different SNS media of private companies. The experimental results show that the proposed Na¨ıve Bayes classifier model has good performance for different SNS media, and a semi supervised learning effectively works for the data consisting of long comments. In addition, the proposed method is applied to detect flaming incidents, and we show that it is successfully detected.

Tweet sentiment analysis is an important research topic. An accurate and timely analysis report could give good indications on the general public's opinions. After reviewing the current research, we identify the need of effective and efficient methods to conduct tweet sentiment analysis [7]. This paper aims to achieve a high level of performance for classifying tweets with sentiment information. We propose a feasible solution which improves the level of accuracy with good time efficiency. Specifically, we develop a novel feature combination scheme which utilizes the sentiment lexicons and the extracted tweet unigrams of high information gain. We evaluate the performance of six popular machine learning classifiers among which the Naïve Bayes Multinomial (NBM) classifier achieves the accuracy rate of 84.60% and takes a few minutes to complete classifying thousands of tweets.

The data potential of Twitter is a powerful resource for data mining exploration. This research aims to pull the traffic information in Jakarta from Twitter. The first output is to develop a web application that can display Jakarta's traffic situation in real time. The process include filtering and tokenizing to get the traffic jam's location and direction to be displayed on Google Map [8]. The second output is to develop a predictive analysis system to oversee Jakarta traffic pattern in a certain period of time using Naïve Bayes Classifier

Text classification is one of the key methods used in text mining. Generally, traditional classification algorithms from machine learning field are used in text classification. These algorithms are primarily designed for structured data. In this paper, we propose a new classifier for textual data, called Supervised Meaning Classifier (SMC). The new SMC classifier uses meaning measure, which is based on Helmholtz principle from

Gestalt Theory [9]. In SMC, meaningfulness of terms in the context of classes are calculated and used for classification of a document. Experiment results show that new SMC classifier outperforms traditional classifiers of Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) especially when the training data limited.

Sentiment analysis and text summarization has evoke the interest of many scientists and researchers in last few years, since the textual data has become useful for many real world applications and problems. Sentiment analysis is a machine learning approach in which machine learns and analyze the sentiments, emotions etc about some text data like reviews about movies or products. These reviews are increasing day by day, due to which summarization of reviews comes in role where summarized form of text in needed, which provides useful information from the large number of reviews[10]. It is very difficult for a human being to extract useful data or summarize it from the very large document. In Text summarization, importance of sentences is decided based on linguistic features of sentences. This paper provides the comprehensive overview of recent and past research on sentiment analysis and text summarization and provides excellent research queries and approaches for future aspects.

To mine the opinion on the web, it is essential to perform a well defined task, which helps us to retrieve the information from the available data on the web. A discussion is started with the introduction on sentiment analysis, which gives us a insight into sentiment analysis [11]. The detail discussion on various methods proposed by different researchers is also presented. Different types of sentiment analysis techniques give a research direction in different directions. Finally a method is proposed based on the naïve bayes classifier.

The increasing use of smart phones to access social media platforms opens a new wave of applications that explore sentiment analysis in the mobile environment. However, there are various existing sentiment analysis methods and it is unclear which of them are deployable in the mobile environment [12]. This provides the first of a kind study in which we compare the performance of 17 sentence-level sentiment analysis methods in the mobile environment. To do that, we adapted these sentence level methods to run on Android OS and then we measure their performance in terms of memory usage, CPU usage, and battery consumption. The findings unveil sentence-level methods that require almost no adaptations and run relatively fast as well as methods that could not be deployed due to excessive use of memory.

IV. System Design:

A. System Architecture Diagram:

An Architecture diagram is drawn for the project which is a pictorial representation of a system, in which principal parts or functions are represented by blocks connected by lines that show relationships of the blocks.

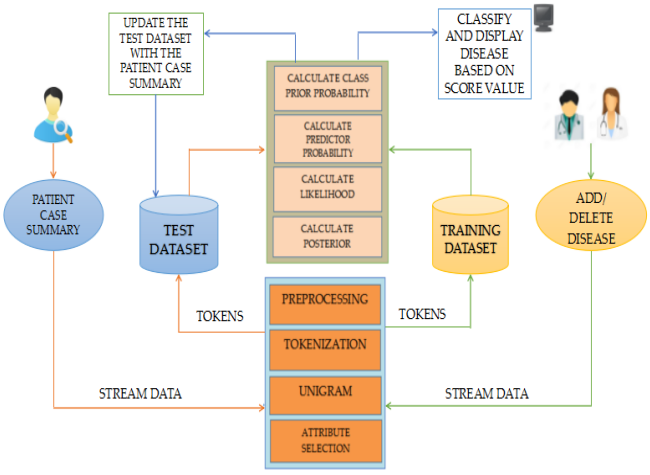


Fig. 1:

B. Existing System:

In Existing system, the single data mining technique is used to diagnose the diseases. There is no previous research that identifies which data mining technique can provide more reliable accuracy in identifying suitable treatment for category of diseases to the patients. The system is not fully automated, it needs doctors for full diagnosis for serious patients.

Practical use of health care database systems and knowledge discovery is difficult in disease diagnosis. Hospitals do not provide the same quality of service even though they provide the same type of service. It takes more time consumption for practical use of health care database systems.

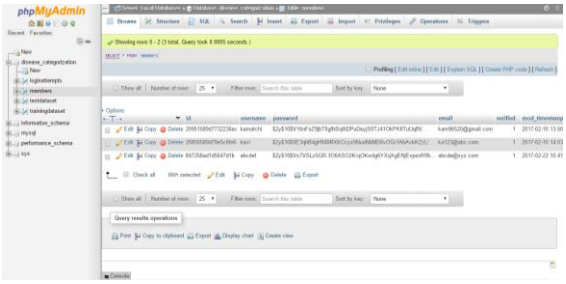
C. Proposed System:

In Proposed System, we are applying hybrid data mining techniques in identifying diseases of the patients. This system can be used by all patients or their family members who need help in emergency. Apply hybrid data mining techniques to the disease diagnosis benchmark dataset to establish baseline accuracy for each single data mining technique in the diagnosis of disease in patients.

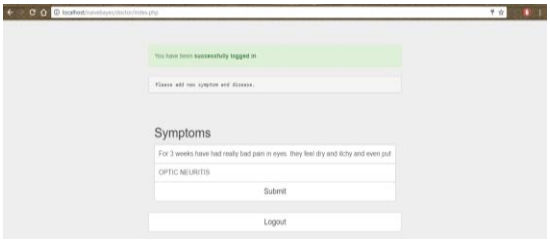
Apply the same hybrid data mining techniques used in disease diagnosis to dataset to investigate if single data mining techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis. Apply the hybrid data mining techniques used in disease diagnosis to data set to investigate if hybrid data mining techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis.

V. Results:

Members Dataset:

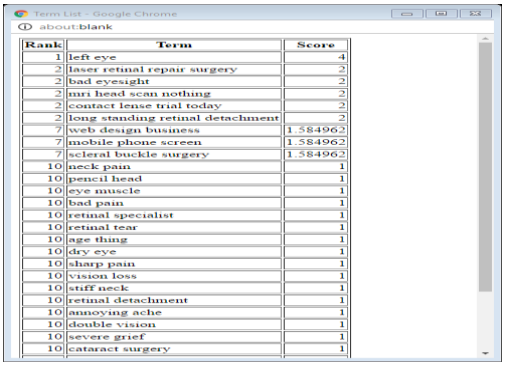


Adding Disease:



Tokenization:

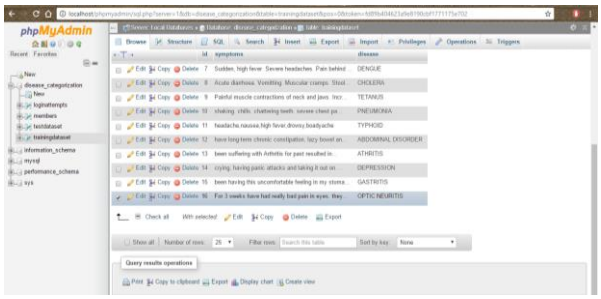
Tokenization Result:



Rank	Term	Score
1	left eye	4
2	laser retinal repair surgery	2
2	bad eyesight	2
2	mri head scan nothing	2
2	contact lense trial today	2
2	long standing retinal detachment	2
7	web design business	1.584962
7	mobile phone screen	1.584962
7	scleral buckle surgery	1.584962
10	neck pain	1
10	pencil head	1
10	eye muscle	1
10	bad pain	1
10	retinal specialist	1
10	retinal tear	1
10	age thing	1
10	dry eye	1
10	sharp pain	1
10	vision loss	1
10	stiff neck	1
10	retinal detachment	1
10	annoying ache	1
10	double vision	1
10	severe grief	1
10	cataract surgery	1

\

Updated Training Dataset:

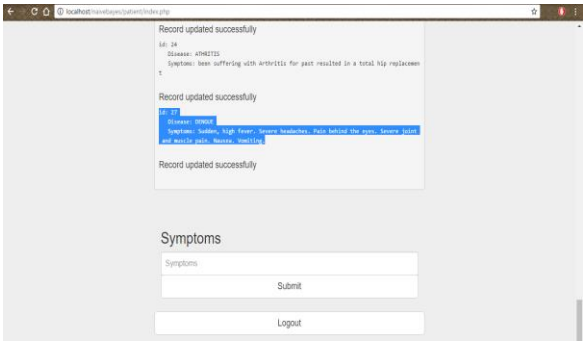


Tokenization Result:

about:blank

Rank	Term	Score
1	fever severe headache pain	2
2	muscle pain nausea	1.584962
2	eye severe joint	1.584962

Displaying Result:



Conclusion:

This project aims to build a system that allows users to get instant guidance on their health issues through an intelligent health care system online. The system is fit with various symptoms and the disease/illness associated with those symptoms. Since Naive Bayes Classifier is based on Probability Theory, the prediction of the disease is made accurately. Unlike the previous system, which is applicable only for single disease classification, this system is applicable for multiple disease classification.

REFERENCES

[1] Bo Tang, Steven Kay and Haibo He, 2016. "Toward Optimal Feature Selection in Naive Bayes for Text Categorization", IEEE transactions on knowledge and data engineering, 28: 9.
[2] Xiaojiang Lei, Xueming Qian and Guoshuai Zhao, 2016. "Rating Prediction based on Social Sentiment from Textual Reviews",IEEE Transactions on Multimedia, Manuscript, 18: 9.
[3] William, B. Claster, Hung Dinh, Malcolm Cooper, 2010. "Naive Bayes and Unsupervised Artificial Neural Nets for Caneun Tourism Social Media Data Analysis".
[4] Aqsath Rasyid Naradhipa, Ayu Purwarianti, 2013. "Sentiment Classification for Indonesian Message in Social Media",2012.Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen and Genshe Chen, "Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier".
[5] Shun Yoshida, Jun Kitazono, Seiichi Ozawa, Takahiro Sugawara, Tatsuya Haga and Shogo Nakamura, 2014. "Sentiment Analysis for Various SNS Media Using Naive Bayes Classifier and Its Application to Flaming Detection".
[6] Ang Yang, Jun Zhang, Lei Pan and Yang Xiang, 2015. "Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination".
[7] Gigih Rezki Septianto, Firman Fakhri Mukti,Muhammad Nasrun S.si., MT, Alfian Akbar Gozali, ST, MT, 2015. "Jakarta Congestion Mapping And Classification From Twitter Data Extraction Using Tokenization And Naïve Bayes Classifier".
[8] A Novel Classifier Based on Meaning for Text Classification,Murat Can Ganiz, Melike Tutkan, Selim Akyokuş, 2015.

[9] Sentiment Analysis and Text Summarization of Online Reviews: A Survey,Pankaj Gupta, Ritu Tiwari and Nirmal Robert, 2016. A Brief Review on Sentiment Analysis, Neha Raghuvanshi, Prof. J.M. Patil.
[10] Towards Sentiment Analysis for Mobile Devices, Johnnatan Messias, Joao P. Diniz, Elias Soares, Miller Ferreira,Matheus Araujo, Lucas Bastos, Manoel Miranda, Fabricio Benevenuto, 2016
[11] Lam, W., M. Ruiz and P. Srinivasan, 1999. "Automatic text ategorizationand its application to text retrieval," IEEE Trans. Knowl. Data Eng., 11(6): 865-879.
[12] Sebastiani, F., 2002. "Machine learning in automated text categorization," ACM Comput. Surveys, 34(1): 1-47.
[13] Al-Mubaid, H. and S. Umair, 2006. "A new text categorization technique using distributional clustering and learning logic," IEEE Trans.Knowl. Data Eng., 18(9): 1156-1165.
[14] Aphinyanaphongs, Y., L.D. Fu, Z. Li, E.R. Peskin, E. Efstathiadis, C.F. Aliferis and A. Statnikov, 2014. "A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization," J. Assoc. Inform. Sci. Technol., 65(10): 1964-1987.
[15] Liu, H. and L. Yu, 2005. "Toward integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowl. Data Eng., 17(4): 491-502.
[16] Salton, G. and C. Buckley, 1988. "Term-weighting approaches in automatic text retrieval," Inform. Process. Manage., 24(5): 513-523.
[17] G€overt, N., M. Lalmas and N. Fuhr, 1999. "A probabilistic descriptionoriented approach for categorizing web documents," in Proc. Int.Conf. Inform. Knowl. Manage., pp: 475-482.
[18] Mnih, A. and G.E. Hinton, 2009. "A scalable hierarchical distributed language model," in Proc. Adv. Neural Inform. Process. Syst., pp: 1081-1088.
[19] Turian, J., L. Ratinov and Y. Bengio, 2010. "Word representations: A simple and general method for semi-supervised learning," in Proc. 48th Annu. Meeeting Assoc. Comput. Linguistics, pp: 384-394.
[20] Joachims, 1998. "Text categorization with support vector Machines : Learning with many relevant features," in Proc. Eur. Conf. Mach.Learn., pp: 137-142.
[21] Lam, W., M. Ruiz and P. Srinivasan, 1999. "Automatic text categorizationand its application to text retrieval," IEEE Trans. Knowl. Data Eng., 11(6): 865-879.
[22] Sebastiani, F., 2002. "Machine learning in automated text categorization,"ACM Comput. Surveys, 34(1): 1-47.
[23] Al-Mubaid, H. and S. Umair, 2006. "A new text categorization technique using distributional clustering and learning logic," IEEE Trans.Knowl. Data Eng., 18(9): 1156-1165.
[24] Aphinyanaphongs, Y., L.D. Fu, Z. Li, E.R. Peskin, E. Efstathiadis, C.F. Aliferis and A. Statnikov, 2014. "A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization," J. Assoc. Inform. Sci. Technol., 65(10): 1964-1987.
[25] Liu, H. and L. Yu, 2005. "Toward integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowl. Data Eng.,17(4): 491-502.
[26] Salton, G. and C. Buckley, 1988. "Term-weighting approaches in automatic text retrieval," Inform. Process. Manage., 24(5): 513-523.
[27] G€overt, N., M. Lalmas and N. Fuhr, 1999. "A probabilistic riptionoriented approach for categorizing web documents," in Proc. Int.Conf. Inform. Knowl. Manage., pp: 475-482.
[28] Mnih, A. and G.E. Hinton, 2009. "A scalable hierarchical distributed language model," in Proc. Adv. Neural Inform. Process. Syst., pp: 1081-1088.