UGC QUALITY EVALUATION BASED ON META-LEARNING AND CONTENT FEATURE ANALYSIS

Xiaoyue Cong[1], Lei Li[2]

[1]Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Beijing University of Posts and Telecommunications, Beijing 100876, China
cxy0105@bupt.edu.cn, leili@bupt.edu.cn

Abstract: With the fast development of Social Networking Services, there has been increasingly vast amount of information published by massive network users. Given this information explosion, how to analyze the quality of User Generated Contents (UGC) automatically becomes a challenging task for researchers. To solve the problem, we need to build an effective UGC quality evaluation system. In the light of our experience, we believe that the textual content of UGC is the key factor for its quality. Hence, we focus on textual content based quality evaluation and classification instead of using UGC publishing related data, such as times being commented and forwarded in this paper. We extract various features of the textual contents based on natural language processing technologies firstly, such as word segmentation, keywords, topic model, sentence parsing, distributed word representation etc. Secondly, we build several base-learning classifiers with different features and different machine learning algorithms to assign UGC contents with four different quality labels. Then, we create the global meta-learning model based on these base classifiers to generate the final quality labels for UGC contents. We have also implemented a series of experiments based on realistic data collected from Tianya Forum and use 10-fold cross-validation to test the model. Results have shown that our proposed meta-learning model performs much better.

Keywords: User Generated Contents (UGC); meta-learning; feature analysis; quality evaluation; multiple classifier fusion

1 Introduction

Different from the leading model of the age of Web1.0, in the Web2.0 era, more and more users not only can get the information from several portal websites, but also can create personalized contents themselves, which is called User Generated Contents (UGC). As the core content in web2.0, UGC implies a high value in various aspects such as economy, politics, culture and so on. Therefore, problems associated with UGC have become the hot topics of research at home and abroad recently.

Different from traditional network data or user log, UGC can be created and published at discretion by everyone. Uneven quality, low value density, high uncertainty and so on are all the specific features of UGC, which also have become basic problems in the analysis of UGC. Therefore, it is necessary for us to assess the quality of UGC itself reasonably.

We can evaluate the UGC quality from many aspects, such as UGC length, the number of browsing, propagation speed, the number of replying, etc. These data are easy to acquire, but they are unreliable indeed. The intervention of human factors such as malicious speculation, advertising and ghostwriters, can create a similar numerical scale but with low quality content. In these circumstances, the simple numerical comparison method is no longer applicable. Therefore, we need a more reliable method to assess the quality of UGC. In this paper, we believe that the textual content of UGC is the key factor for its quality. So we extract multiple features from UGC texts, and form a quality classifier by meta-learning model so as to evaluate the quality of UGC robustly.

2 Related work

2.1 Content quality analysis

Text quality evaluation has always existed in text processing, information retrieval, data mining, natural language processing and other related areas. Up to now, it has made some good achievements for the specific kind of traditional text, essays.

As early as 1966, Page et al. began to study automated essay scoring project. They build the first widely used system called Project Essay Grading(PEG) [1]. After that, there have existed several kinds of automatic scoring systems.

Intelligent Essay Assessor (IEA) [2] is based on Latent Semantic Analysis (LSA). Each document is regarded as a vector, and all of them (including reference essay) constitute a matrix. Then, reduce dimensions of the matrix by singular value decomposition (SVD) and calculate the cosine similarity between essays and the first n reference essays. Finally, calculate the score by a weighted average of the cosine similarity.

E-rater [3], developed by Educational Testing Services (ETS) in the last of 1990s, has been currently put into use for essay scoring in the Graduate Management Admissions Test (GMAT) and the Test of English as a Foreign Language (TOEFL). The E-rater system extracts both shallow and deep text features, and then makes a linear regression equation to score essays.

978-1-5090-1246-6/16/$31.00 ©2016 IEEE

There are also other essay scoring systems, such as Bayesian Essay Test Scoring sYstem (BETSY) [4] and so on. On the basis of the research literatures [5,6], the common methods of essay scoring system include: Regression Model, Vector Space Model (VSM), Bayesian Classification, LSA, Feature Selection, etc. Yet there is rare similar work for UGC.

2.2 Meta-learning

Meta-learning is a subfield of Machine Learning where automatic learning algorithms are applied on meta-data about machine learning experiments. The main goal is to use such meta-data to understand how automatic learning can become flexible in solving different kinds of learning problems, hence to improve the performance of existing learning algorithms.

Mario et al. build a model that can be used to predict algorithm performance when a new optimization problem is presented. [7] Taciana et al. combine meta-learning and search algorithms to deal with the problem of SVM parameter selection. Meta-learning is employed to recommend SVM parameter values based on parameter configurations that have been successfully adopted in previous similar problems when given a new problem to be solved in this combination. [8] Haochang Wang et al. present a novel meta-learning based classifier ensemble model. Four classifiers i.e. Generalized Winnow, support vector machine(SVM), conditional random fields(CRF) and maximum entropy are combined using two different meta-learning strategies. Experimental results show that the novel classifier is obviously superior to the individual classifier based method and superior to the arbitration rule based ensemble method. [9]

In a word, meta-learning based classifier can merge each individual classifier's performance and achieve a better result.

3 Machine learning algorithm and fusion

Generally speaking, each kind of machine learning algorithm has its advantages and disadvantages for various kinds of applications. Similarly, different kinds of features have different contributions for various types of data. As to the task of classification, we can look on these different algorithms and features as base classifiers. Just like a lot of researchers, we believe that fusion of multiple base classifiers with different machine learning algorithms and features has the potential to learn from other's strong points to make up one's deficiencies and thus improve the overall performance. This is just the central idea of this paper.

Basically, it depends on the performance and independence of the base classifiers whether multiple classifier fusion is effective or not. In order to keep the independence of errors produced by the base classifiers, it should meet one of the two criteria: one is different feature description with the same machine learning algorithm, the other is using different classification algorithms for the base classifiers.

In this paper, we adopt two different base machine learning algorithms: SVM and CRF, which are among the best ones in most natural language processing tasks.

3.1 SVM

SVM is a kind of machine learning algorithm based on kernel function method. Its main idea is mapping the data which can't be linearly classified in low dimensional space into high dimensional space by kernel function, and then, find the best hyperplane to divide the data into two categories. Please refer to [10] for more details.

3.2 CRF

The idea of CRF model comes from the maximum entropy model. We can put the CRF as an undirected statistics graph model and markov random field. It is suitable for serialized data task of natural language processing field. [11]

4 Content feature selection

4.1 Keyword

After Chinese word segmentation and stop words removing, we choose two methods to select keywords for content representation respectively. One is the Chi-square test (CHI) [12], the other is the information gain (IG) [13].

Chi-square test is a commonly used statistical method of inspecting the independency between two variables. The common practice is to assume that the two variables are independent (null hypothesis), then, calculate the difference between the theoretical value and the practical value.

Table I shows some of the selected keywords by Chi-square test for different UGC with four different quality labels, which are Grade 0, 1, 2 and 3. More details of these labels are given in Table III.

Table I Some selected keywords by Chi-square test

Grade 0		Grade 1	
keyword	**CHI**	**keyword**	**CHI**
日	39129.77	中国	240776.8
月	27083.11	人	182722.6
Grade 2		**Grade 3**	
keyword	**CHI**	**keyword**	**CHI**
weibo	100465.2	山石	31245.7
要闻	97454.9	屏	29866.8

After we get the keywords for different UGC with four different quality labels, we put these keywords in a set in order to form the feature set.

The information gain is a measure of the difference between two probability distributions P and Q. It is not symmetric in P and Q. In applications, P typically represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while Q typically represents a theory, model, description, or approximation of P.

Table II shows some selected keywords by information gain for the classification system. The IG of a word expresses the importance of the word for the classification system not for any quality label of the UGC.

Table II Some selected keywords by information gain

keyword	IG	keyword	IG
中国	0.113844	官员	0.076107
不是	0.090332	人	0.075696

4.2 Dependency grammar

Dependency grammar (DG) is a class of modern syntactic theories that are all based on the dependency relation and that can be traced back primarily to the work of Lucien Tesnière. Dependency is the notion that linguistic units, e.g. words, are connected to each other by directed links. The verb is taken to be the structural center of clause structure. All other words are either directly or indirectly connected to the verb in terms of the directed links, which are called dependencies. DGs are distinct from phrase structure grammars since DGs lack phrasal nodes although they acknowledge phrases. Structure is determined by the relation between a word and its dependents.

We get the dependency grammar of sentences using a toolkit named Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml). Then, we take all of the local syntactic structure and dependence of each post as features to describe the post.

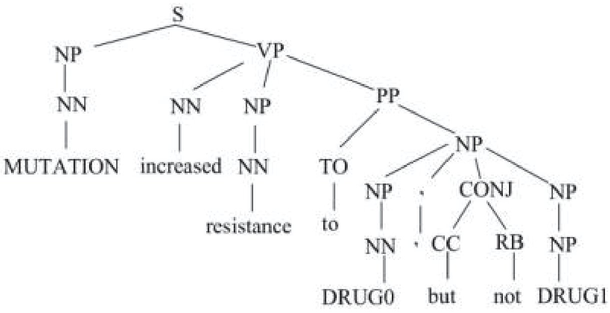

Figure 1 An example of syntactic structure

In Figure 1, for example, we take structures like NP-NN or VP-NN-NP-PP as a feature.

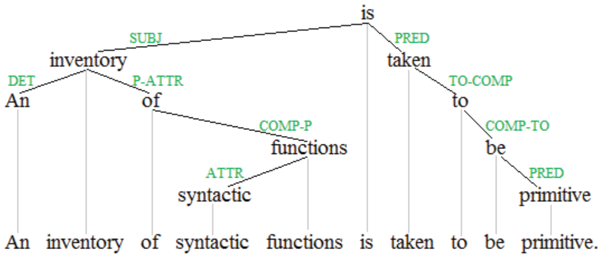

Figure 2 An example of dependence

In Figure 2, for example, we take SUBJ et al. as a feature.

4.3 Latent dirichlet allocation topic model

Topic model is a generative model for documents; it specifies a simple probabilistic procedure by which documents can be generated. Latent Dirichlet Allocation (LDA) [12,14] is one of the most popular topic models. LDA assigns a discrete latent model to words and let each document maintain a random variable, indicating its probabilities of belonging to each topic.

In this paper, we use LDA topic model to generate the topics of the four types' posts.

4.4 Doc2vec

In 2014, based on the efficient estimation of word representations in vector space "Word2vec" [15], Mikolov et al. proposed distributed representations of sentences and documents, named "Doc2vec" [16]. Doc2vec can combine words with the full context, semantics, grammar and emotional information quite well, therefore, this paper introduces this method as an expression of the text feature. We will explore its effect in text quality evaluation.

5 Meta-learning model

We use each of the above features with SVM and CRF algorithms respectively to train various base classifiers. Then we fuse these base classifiers according to meta-learning model so as to obtain an improved higher level meta classifier. The strategies we used for fusion from base level to higher level are: stacked generalization, D-S theory of evidence and voting. Finally, we will compare the results of the three methods. Please refer to [17] for more details of D-S theory of evidence. Voting is a very simple scheme in which the final results are those with more votes from the base classifiers.

Figure 3 shows the flow diagram of stacked generalization in meta-learning framework.

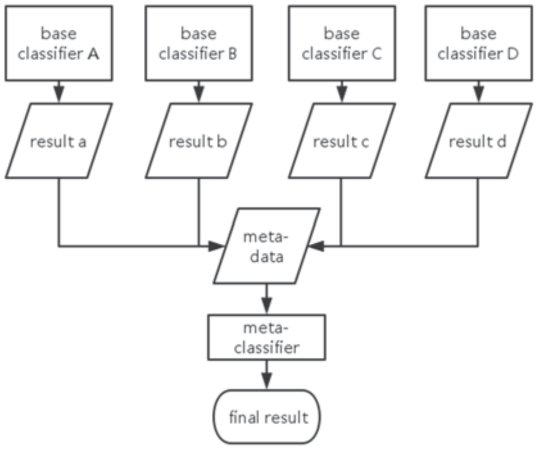

Figure 3 Flow diagram of meta-learning

6 Experiments and analysis

6.1 Data sources

We collected experimental data from Tianya Forum (http://bbs.tianya.cn) using our own web crawler. Finally, we get 1018750 posts from the Forum. In Tianya Forum, each post has a grade label from 0 to 3. Different grade marked by board masters represents different quality. So, we consider the UGC quality evaluation as a classification problem. The meanings of the four labels are shown in Table III. We will see, the difference between grade 0 and grade 2 is just the key points.

Table III The meanings of the four labels

grade	meaning	quality
0	black humor	between grade 1 and 3
1	normal post	worst
2	suggested reading	between grade 1 and 3
3	original boutique	best

Table IV The numbers of the four kinds of posts

grade	number
0	2475
1	1012198
2	770
3	3307

Table IV shows the numbers of the four kinds of posts in our collected data. In order to keep the balance of the data, we finally choose 2960 posts (there are 740 posts of each kind) as the experimental data after data pretreatment. We only use the main content of each post for classification.

6.2 Experiments of base classifiers

In this paper, we adopt 10-fold cross-validation to test the performance of each base classifier. Then, as for each classifier, we will get a predicted value for each post of the experimental data. We can calculate the classifier performance based on the prediction.

In order to get the important degree sorting of the syntactic structures and dependencies, we use the information gain and Chi-square test to calculate it. And we adopt Sequential Forward Selection (SFS) method to confirm the number of keywords, the number of dependencies, the number of syntactic structures and the number of topics for LDA.

It is worth mentioning that we choose the zero (0) to represent the absence of a keyword and we do an experiment about how to represent the presence of a keyword. The first method is the one (1) to represent the presence of a keyword. The second method is using the sentiment score of the keyword from a sentiment dictionary, while the last method is the term frequency of the keyword in the post. Table V shows the result of SVM of the three methods.

Table V Results of the three methods

keyword by CHI	cross validation accuracy
0,1	55.61%
0,sentiment score	48.99%
0, term frequency	44.39%

We can see that the best method is the first method. So we choose this method to represent the presence of a keyword, a syntactic structure or a dependency. Table VI shows the 14 base classifiers to be fused and their performance.

Table VI Results of base classifiers

feature	SVM	CRF
Keyword(IG)	54.93%	55.78%
Keyword(CHI)	55.61%	55.64%
LDA	39.53%	46.01%
Syntactic structure(IG)	51.82%	51.42%
Syntactic structure(CHI)	51.99%	51.79%
Dependency(IG)	43.85%	43.41%
Dependency(CHI)	44.29%	43.95%

In Table VI, we can see that this method of UGC quality evaluation is feasible with the best result reaching 55.64%. But different features have influence on the result of classification deeply, different machine learning algorithms also have an effect on the result.

As to the Doc2vec, we tested it separately. We use a Java implementation of doc2vec in ICML'14 (https://github.com/yao8839836/doc2vec_java) to get the vector of each document. The document vector dimension is 200 by default. Firstly, we look on the 2960 posts as a corpus. By 10-fold cross-validation of the SVM, we get an amazing effect. The cross validation accuracy reached 88.277%, much higher than the result of any of these features. But as we know, Doc2vec cannot generate the document vectors incrementally and we think that size of the data for doc2vec is too small. So we add the training data of doc2vec to 1GB and after the progress, we take out the document vectors of the 2960 posts to do the 10-fold cross-validation of the SVM. This time, the result is not satisfactory. The cross validation accuracy is only 35.777%. Finally, we don’t choose this feature for fusion.

6.3 Experiments of multiple classifier fusion

After 10-fold cross-validation, for each posts we have a list of prediction labels from the base classifiers. So, we have a result file just like Table VII for each base classifier. The first column is line number representing each post, the second column is the golden grade labels of each posts, the third is column of the prediction labels, and the subsequent four columns are the prediction probability of four labels. Then, we can fuse all base classifiers using these results and test the fusion methods in our meta-learning framework.

Table VII Format of the result file of each base classifier

line	grade	labels	0	…	3
0	0	2	0.363	···	0.100
1	0	1	0.376	···	0.088
…	…	…	…	…	…
2959	3	1	0.240	···	0.114

We have tried to use all the features to train only one base classifier. But the result is not the best among the results in Table VI. So, we try to fuse the base classifiers using three different strategies for meta-learning, which are stacked generalization, D-S theory of evidence and voting.

The result of the three methods is shown in Table VIII.

Table VIII Result of classifiers' fusion by three methods

method	accuracy
Stacked generalization	62.84%
D-S theory of evidence	59.93%
Voting	56.05%

In Table VIII, the accuracy is the 10-fold cross validation accuracy. We can see that the stacked generalization is the best method among the three ways of classifiers' fusion. Table IX shows the detailed result of each label from the meta-learning classifier with fusion strategy of stacked generalization.

Table IX Detailed result of classifiers' fusion by stacked generalization

grade	precision	recall	F1-score
0	57.97%	55.28%	0.566
1	76.35%	78.80%	0.776
2	54.87%	58.42%	0.566
3	62.16%	59.59%	0.608
Entirety	62.84%	62.84%	0.628

From Table VIII and Table IX, we can find out that the result of stacked generalization classifier can obtain more than 7% of the best base classifier.

To sum up, all the above experimental results show that UGC quality evaluation based on content feature analysis in meta-learning framework is feasible and the influence of the feature selection for system performance is also very important. In addition, stacked generalization is an effective fusion strategy for meta-learning framework.

7 Conclusions

In this paper, a new classifier fusion model based on the meta-learning strategy is proposed and applied to the UGC content quality evaluation. This method takes advantage of the complementarity and correlation between different classifiers. Experiments have shown that its performance is superior to all the base classifier systems. At the same time, the features selected in allusion to the UGC quality evaluation are effective. We will continue to work on more kinds of UGC in the future.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 91546121, 61202247, 71231002 and 61472046; EU FP7 IRSES MobileCloud Project (Grant No. 612212); the 111 Project of China under Grant B08004; Engineering Research Center of Information Networks, Ministry of Education; the project of Beijing Institute of Science and Technology Information; the project of CapInfo Company Limited.

References

[1] Page, Batten E. Computer Grading of Student Prose, Using Modern Concepts and Software[J]. Journal of Experimental Education, 1994, 62(2):127-142.

[2] Landauer T K, Foltz P W, Laham D. The intelligent essay assessor[J]. Intelligent Systems IEEE, 2000, 15(5):27-31.

[3] Attali Y, Burstein J. Automated Essay Scoring With e-rater® V.2[J]. Journal of Technology Learning & Assessment, 2004, 4(3):i–21.

[4] Rudner L M, Liang T. Automated essay scoring using Bayes' theorem[J]. National Council on Measurement in Education New Orleans La, 2002, 1(2):3--21.

[5] Huang Z, Xie J, Xun E, et al. Study of feature selection in HSK automated essay scoring[J]. Computer Engineering & Applications, 2014.

[6] Shermis M D. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration[J]. Assessing Writing, 2014, 20:53-76.

[7] Muñoz M A, Kirley M, Halgamuge S K. A meta-learning prediction model of algorithm performance for continuous optimization problems[M]//Parallel Problem Solving from Nature-PPSN XII. Springer Berlin Heidelberg, 2012: 226-235.

[8] Gomes T A F, Prudêncio R B C, Soares C, et al. Combining Meta-Learning and Search Techniques to Select Parameters for Support Vector Machines[J]. Neurocomputing, 2012, 75(1):3–13.

[9] Wang H C, Zhao T J, Zheng D Q, et al. Meta-learning based classifier ensemble strategy and its application[J]. Journal on Communications, 2007.

[10] Boser B E. A training algorithm for optimal margin classifiers[C]// PROCEEDINGS OF THE 5TH ANNUAL ACM WORKSHOP ON COMPUTATIONAL LEARNING THEORY. 1992:144--152.

[11] John D. Lafferty, Andrew McCallum, Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models For Segmenting And Labeling Sequence Data[C]// 2001:282--289.

[12] Zheng C, Xiong D K, Liu Q Q. The Short Text Classification Method Based on CHI Feature Selection and LDA Topic Model[J]. Computer Knowledge & Technology, 2014.

[13] Dong W, Liu X, Hong N I. Adaptive feature selection method based on information gain[J]. Computer Engineering & Design, 2014.

[14] Jin-Qun H E, Liu P J. The documents classification algorithm based on LDA[J]. Journal of Tianjin University of Technology, 2014.

[15] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[16] Le Q V, Mikolov T. Distributed Representations of Sentences and Documents[J]. Eprint Arxiv, 2014, 4:1188-1196.

[17] Mandler E, Schümann J. Combining the Classification Results of Independent Classifiers Based on the Dempster/Shafer Theory of Evidence[C]// Pattern Recognition and Artificial Intelligence. Towards an Integration. 1988:381-393.