Quality Assessment Method of Web Documents Based on Random Forest

Li He[1,a,*], Li Tang[1,b] and Ning Wang[1,c]

[1]School of Science and Technology, Tianjin University of Finance and Economics, Tianjin 300222, China

[a] renkeheli@163.com, [b] tangli0831@ tjufe.edu.cn, [c] ninglw@163.com

*corresponding author

Keywords: Web document, quality assessment, LDA topic model, random forest.

Abstract: This paper proposes a method based on the method of Random Forest (RF) for better assessing quality of web documents, and formulates a novel quality evaluation index system including features of organization structure, network access and content. In order to extract the content feature of a document, a topic coverage degree calculation model based on LDA is put forward. Finally, it conduct some experiments on two document sets: Wikipedia and Baidu Encyclopedia, and precision rate, recall rate and F-Measure are used to verify the validity of the proposed quality assessment method. Experimental results show that the proposed evaluation index system and the RF-based quality assessment method can achieve good performance and advantages.

1. Introduction

The web document is kind of important data in the era of internet, and it has complex and diverse structure. It is necessary to achieve the quality-based web document management. Data quality assessment is an important part of data management. The essence of data quality assessment is to evaluate the data quality according to the predefined quality dimensions. The typical quality dimensions of structured data mainly include correctness, integrity, stability, consistency and timeliness [1-3]. Due to the openness of the Internet, and the lack of supervision mechanism and the standardization of constraints, there are significant differences in the quality of documents with the same topic. Therefore, in order to effectively manage web documents, and to provide users with high-quality document services, it is necessary to explore more effective methods for the quality assessing of web documents.

The web document is unstructured and open, which makes the quality depending on not only the content, but also its organization structure and the access by internet users and authors. In the aspect of data quality assessment of web documents, many researchers have paid more attention to quality dimensions and evaluation methods. According to the relationship between quality and authors of academic authority, literature [4] put forward an evaluation method based on the edit history and the author academic authority. Literature [5] proposed a quality evaluation method based on the document length. Literature [6] presented a quality evaluation method based on the support vector regression, the assessment indexes included browsing history, text and network features and so on. Literature [7] put forward a method based on the lexical cue model. Literature [8] proposed a new method based on Hidden Markov model according to the document revision cycle. Literature [9] considered that the higher survival rate after many edits, the better of the document quality, and put forward a method based on the author reputation and interactive score. Literature [10] used the meta-learning technology to combine data quality indexes for the semantically associated documents, and proposed an automatic evaluation method based the meta- learning technology.

The above researches are mainly based on two aspects: document access features or content features. In fact, the quality of web documents is related to not only their access characteristics, but also the organizational structure and content characteristics. In this paper, we formulate the quality evaluation index system including six dimensions, such as correctness, availability, integrity, consistency, readability and timeliness. In order to get the content feature, we proposes a topic coverage degree calculation formula based on LDA topic model, and put forward a new quality

evaluation method based on RF. The main contributions of this paper are as follows: (1) content, structure and openness are taken into account in the quality assessment algorithm of web documents; (2) the topic coverage degree calculation model based on LDA topic model is formulated; (3) the RF classifier model is used to assess the web document quality.

The rest of this paper is organized as follows: Sect.2 expounds the RF theory. The quality dimensions and quality evaluation method based on RF for web documents are proposed in Sect.3. Sect.4 discusses the experimental setup and gives the analysis of experimental results. At last, we draw a conclusion in Sect.5.

2. RF Fundamental Theory

RF is proposed by Leo Breiman and Adele Cutler in 2001, it is a kind of classifier combination model which can deal with high dimension and nonlinear samples. RF can be defined as a set of $\{h_1(\boldsymbol{X},\theta_i), 1\le i\le K\}$. Where, K represents the number of decision trees contained in a forest, $\{\theta_i\}$is an independent and identically distributed random vector that determines the growth process of a single decision tree，$\boldsymbol{X}$ is an input vector. RF uses bootstrap re-sampling technique and random feature selection method to establish a decision tree, and then determines the class label of $\boldsymbol{X}$ by decision trees in RF. RF has good robustness to noise and missing data, it has fast learning speed, and can provide the importance analysis of each classification features. Therefore, it has been widely used in classification, prediction and important feature selection fields.

Given the set of classifiers $\{h_1(\boldsymbol{X}), h_2(\boldsymbol{X}),...,h_K(\boldsymbol{X})\}$,and the training set obtained from the random vector $\boldsymbol{X}$ and $\boldsymbol{Y}$, the performance evaluation of the RF classifier is discussed as follows.

2.1 RF Generalization Error[11]

Definition 1 margin function. The margin is used to measure the difference between the average number of votes on the correct class and the average number of votes on the error one. It is defined as Eq. (1):

$$mg(X,Y)=av_K I(h_K(X)=Y)-\max_{j\neq Y} av_K I(h_K(X)=j) \tag{1}$$

In the formula (1), I (.) is the indicator function. The margin measures the extent to which the average number of votes at $\boldsymbol{X}$, $\boldsymbol{Y}$ for the right class exceeds the average vote for any other classes. The larger the margin, the higher the probability of classification accuracy. In RF, $h_i(\boldsymbol{X})=h(\boldsymbol{X}, \theta_i)$, $1\le i\le K$, Therefore, the RF margin function can be shown as Eq. (2):

$$mr(X,Y)=P_\theta(h_K(X)=Y)-\max_{j\neq Y} P_\theta(h_K(X)=j) \tag{2}$$

Definition 2 generalization error. The generalization error of RF is defined as Eq. (3):

$$PE^*=P_{X,Y}(mr(X,Y)<0) \tag{3}$$

In Eq. (3), $P_{X,Y}$ represents the probability in space $\boldsymbol{X}$ and $\boldsymbol{Y}$. The generalization error describes the ability of the trained classifier to be applied to a new data set. The learning ability of RF is stronger when the generalization error is smaller. When mr ($\boldsymbol{X}$, $\boldsymbol{Y}$) <0, it indicates that the test sample is wrong classified by RF. Therefore, PE is actually the probability that RF wrong classify the test sample.

Theorem 1 with the increase of the number of decision trees in RF, for sequences $\{\theta_1,...,\theta_K\}$, PE^* converges to:

$$P_{X,Y}(P_\theta(h(X,\theta)=Y)-\max_{j\neq Y} P_\theta(h(X,\theta)=j)<0) \tag{4}$$

Theorem 1 shows that RF generalization error converges to a limit value, and the over-fitting problem can be avoided.

Define 3 Out Of Bag error (OOB error). In order to construct a RF decision tree, the training set is produced using random sampling with putting back. When the sample size is large enough,

about 37% of samples cannot be extracted, which are called OOB data. The error rate calculated on the OOB data is called OOB error, which is usually used to evaluate the generalization ability of RF.

2.2 Importance Measure of Classification Features

RF provides an important function which evaluates the importance of each feature, and it can sort the contributions of different features in RF classifier.

Definition 4 **the importance of the feature R**. Assume the number of classes in RF is C, $A_{OOB}(R)$ and $A'_{OOB}(R)$ are used to represent the correct rate of OOB data before and after a slight disturbance on the feature R respectively, $I_{OOB}(R)$ is used to denote the importance of R, then we define $I_{OOB}(R)$ *as* Eq. (5).

$$I_{OOB}(R) = \frac{1}{K}\sum_{i=1}^{C}(A_{OOB}(R) - A'_{OOB}(R)) \tag{5}$$

When the random noise is added to *R*, if the classification accuracy of the OOB data is greatly decreased, the feature R has a great impact on the classification results.

3. RF Quality Evaluation Algorithm for Web Documents

Web document quality evaluation is actually a process of document classification based on quality features, the main steps include: (1) defining document quality features; (2) formulating the index system of quality evaluation; (3) constructing RF-based quality assessment algorithm for web documents.

3.1 Quality Feature Extraction of Web Documents

In view of the openness and unstructured characteristics, this paper defined six quality dimensions including timeliness, usability, correctness, completeness, readability and usability. First of all, the interactions between authors and the document will affect the quality; secondly, the content integrity and the topic coverage degree of a document have an important influence on the quality; thirdly, the organizational structure of the document is also influenced the readability. Therefore, in addition to the basic features of a web document, this paper considers four main quality features: organizational structure, network structure, access and content characteristic. Based on these five characteristics and six quality dimensions, we define sixteen assessment quality indexes shown in Table 1 for the web documents.

In Table 1, the number of document length, section number, table and picture number, references and footnotes are used to describe the structure of the document; features such as the number of redirection, in-degree and out-degree are applied to the description of the network structure. The in-degree is used to represent the number of link to the document from the other pages, and the out-degree is used to express the number of link to other pages from the page of the document. The topic coverage degree is used to represent the content feature, and all other features are used to describe the access feature of the document.

The structural features, network features and access features of the document can be extracted directly from the document page, while the topic coverage degree should be calculated based on the document content. In this paper, the LDA topic model is used to extract the document content feature.

LDA topic model is a generative model which can identify the latent topics for large-scale document collections. In LDA topic model, a document can belong to multiple hidden topics with different probabilities, and the greater the probability, the higher the coverage degree of the document and the topic.

Table 1 Quality evaluation indexes and dimensions of web documents

Number	Evaluation index	Quality dimension
1	Setup time	timeliness
2	Document length	Correctness
3	Section number	Readability
4	Sub section number	Readability
5	Number of pictures and tables	Readability
6	Number of citations and	Integrity
7	Number of redirection to the	Usability
8	In-degree	Usability
9	Out-degree	Usability
10	Number of concerned users	Usability
11	Number of languages	Usability
12	Daily revise times	Correctness
13	Total number of authors	Correctness
14	Number of changes within 30	Uniformity
15	Number of different authors	Uniformity
16	Topic coverage degree	Integrity

For a document d, if we use $\boldsymbol{P}=<P_1,P_2,...,P_K>$ to represent the probability distribution of d in K topics, we use *acc_cq* to denote the topic coverage degree of d.

$$acc\_cq = \sqrt{\sum_{i=1\sim K} COT_i \times P_i} \tag{6}$$

In Eq. (6), P_i represents the probability of d belongs to the topic T_i described as the i-th topic, $1\le i\le K$, $0\le P_i\le 1$; COT_i expresses the coverage degree of the keyword feature vector on the topic T_i. We use $f=<w_1,w_2,...,w_l>$ to express the keyword feature vector of the document d, $tf_i=<tw_{i1},tw_{i2},...,tw_{in}>$ to represent the keyword feature vector of topic T_i, and tw_{is} $(0\le s\le n)$ to denote the s-th keyword of T_i, w_k $(1\le k\le l)$ to represent the k-th keyword of the document d, then the calculation method of COT_i is shown as Eq. (7):

$$COT_i = (\sum_{k=1\sim l, s=1\sim n} syn(tw_{is}, w_k)) / NK_i \tag{7}$$

In Eq. (7), $syn(tw_{is},w_k)$ calculates the semantic similarity between keywords w_k and tw_{is}, NK_i represents the keywords contained in T_i.

The coverage degree indicates the degree of correlation between a document and topics. The greater the value of *acc_cq*, the higher relevance degree to the topic.

3.2 Performance Evaluation of the Classifier

In the classification system, Precision (P), Recall (R) and F-measure (F) are three most commonly used classification performance evaluation index. TP is used to denote the number of samples that are correctly assigned to the positive class; TN is used to express the number of samples which are not a case of the positive class and are not assigned to it; FP is used to represent the number of samples which are wrong assigned to the number of the sample; FN is used to denote the number of samples which do not belong to the positive class but are assigned to it. P, R and F are defined as Eqs. (8), (9) and (10):

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$F = \frac{2 * P * R}{P + R} \tag{10}$$

3.3 RF Quality Evaluation Algorithm

After extracting the quality features of web documents, a quality feature set used as the original sample set can be built for RF training. Because dimensions of evaluation indexes are usually different, it is firstly necessary to normalize the original samples before RF training. RF parameters such as the number of decision trees (*Ntree*), the number of features participating in training of decision trees, and node splitting conditions should be also set before.

We denote the normalized data set as *D*, the total number of features as *M*, the number of features used for node splitting as *m* (m<<M), the main steps of the Document Quality Evaluation algorithm based on RF (DQERF)are described in Fig. 1.

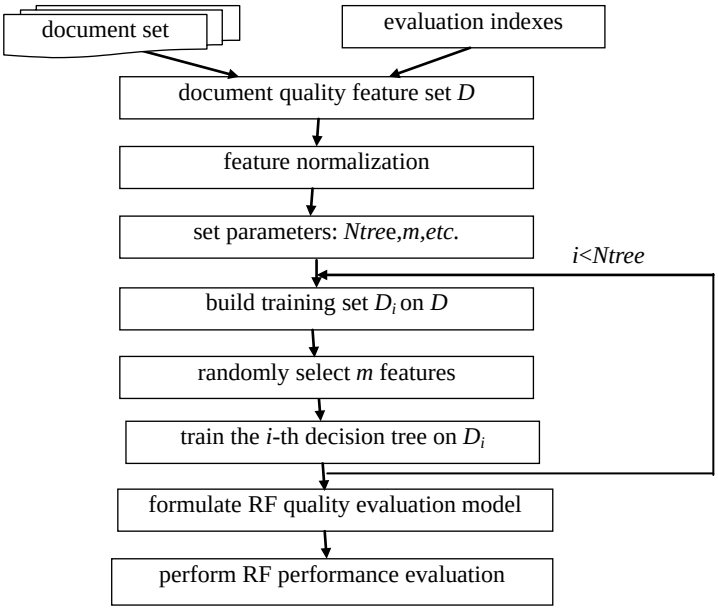


Fig. 1 RF training process of web document quality evaluation

4. Experiment and Result Analysis

Three are two experiment purposes: (1) validating the effectiveness of the quality evaluation index which combines content, network, structure and access features; (2) verifying the availability of DQSRF.

4.1 Experimental Design

Firstly, we randomly select 236 English documents from Wikipedia website belong to the fields of education, engineering, economics and geology; 126 Chinese documents from Baidu Encyclopedia website including economy, nature and technology, and then build English and Chinese document feature data set denoted separately as *D1* and *D2* according to the definition of document quality features described in Table 1. The documents in *D1* and *D2* have been classified into three quality levels: FA (Featured Article), GA (Good Article) and B (Bad articles), which are shown in Table 2.

Table 2 Experiment data sets

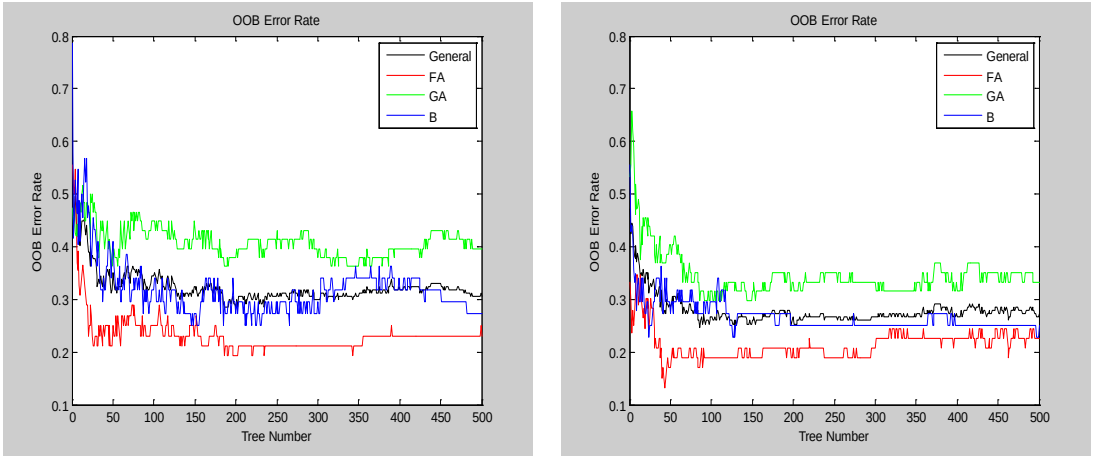
Datasets	FA	GA	B
D1	62	74	69
D2	43	48	35

In order to achieve RF quality prediction, we use 1, 2 and 3 to denote the level of FA, GA and B respectively, the topic number of LDA is set to 3, LDA topic model is generated using the UMASS toolkit mallet [16]. In the training of RF, *m* is set to $\log_2^M$+l, the number of decision trees *Ntree* is set to 500.

4.2 Analysis of Experimental Results

4.2.1 Performance Evaluation of RF Classifier

Because of the random of the training set and the feature selection for node splitting during the RF training process, we repeatedly run the algorithm DQSRF ten times on *D1* and *D2* respectively, and use the average of *P*, *R* and *F* in ten times as the performance evaluation results of DQERF. Fig. 2 shows the results of OOB error rate when the RF is trained on the *D1* and *D2*.



(a) OOB error rate on D1 (b) OOB error rate on D2

Fig. 2 RF OOB error rate during RF training

In Fig. 2, (a) describes the OOB error rate when the error rate is 0.176863 on *D1*; (b) describes the OOB error when the error rate is 0.254902 on *D2*. From Fig.2, the average OOB error rate of *D1* is 0.2453, the average OOB error rate of *D2* is 0.2978, and the overall correct rate exceeds 70%. Because the documents labelled as FA and B can be more easily differentiated, we find that the OOB error rate of DQERF on the FA and B samples is lower than GA documents. In order to further illustrate the effectiveness of DQERF algorithm, Table 3 shows the average of *P*, *R* and *F* in ten times on *D1* and *D2*.

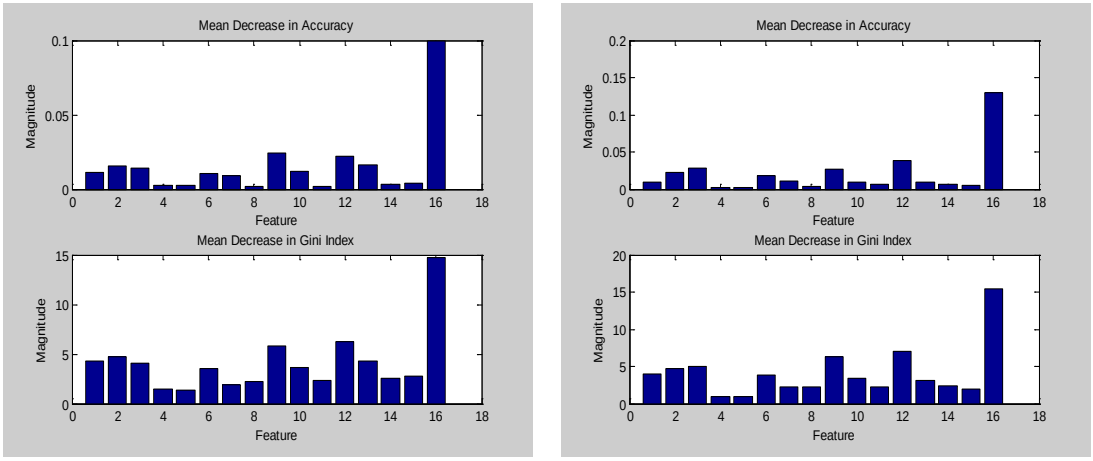
Table 3 Performance of DQSRF

performance	*D1*			*D2*		
indexes	FA	GA	B	FA	GA	B
P	0.824	0.694	0.806	0.808	0.674	0.79
R	0.834	0.664	0.832	0.814	0.692	0.762
F	0.829	0.679	0.819	0.811	0.683	0.776

From Table 3, we find that the DQSRF algorithm can obtain good assessment effect on the quality evaluation of web documents. The overall accuracy rates on *D1* and *D2* are separately 79% and 77%, the average recall in *D1* and *D2* are separately 77% and 75%, and the F-measure index reaches more than 80% in the class FA samples. These results show that the DQSRF algorithm is effective on the quality assessment of web documents.

4.2.2 Importance Analysis of Quality Features

RF provides the importance evaluation of features. In RF training, the Gini index value of a feature indicates its contribution to the classification result. In order to verify the content quality index based on LDA topic model proposed in this paper, Fig. 3 shows the Gini index changes of each feature on *D1* and *D2*.



(a) The feature Gini index of *D1* (b) The feature Gini index of *D2*

Fig. 3 Comparison of the feature Gini indexes of *D1* and *D2*

In Fig. 3, (a) described the influence of each feature on *D1* when the error rate is 0.176863; (b) depicts the influence of each feature on *D2* when the error rate is 0.254902. From Fig.3, we find that the importance of each feature on *D1* and *D2* is basically the same, and the sixteenth feature has the most important effect on the accuracy of RF-based quality evaluation. Therefore, we can conclude that the content quality based on LDA topic model has more obvious influence on the average accuracy of DQSRF algorithm.

5. Conclusion

The quality evaluation of web documents is an important basis for document filtering and intelligent recommendation in the era of network, and it is also an important part of the quality management of web documents. In view of network and unstructured features of web documents, this paper formulates a quality evaluation index system based on organization structure, network structure, access and content features. In order to extract the content feature, the coverage degree based on the LDA topic model is built. Afterwards, a new quality evaluation algorithm based on RF classifier model is implemented. For evaluating the effectiveness and advantages of the proposed DQERF algorithm, we perform some experiments on the English document set and Chinese document set respectively. Experimental results show that DQERF algorithm can achieve good evaluation performance in precision, recall rate and F-measure. Especially, it can obtain a very good precision and recall rate for the quality evaluation of good and poor documents. Meanwhile, the experimental results also show that the content feature based on the coverage degree has important contribution to the quality evaluation results. In the future, we will continue to explore the quality evaluation method of web documents based on quality feature optimization.

6. Acknowledgment

This work was financially supported by the Tianjin Nature Science Fund (No.15JCYBJC16000).

References

[1] Batini C, Cappiello C, Francalanci C, et al. Methodologies for data quality assessment and improvement [J]. ACM Computing Surveys (CSUR), 2009, 41(3):16.

[2] Kwon O, Lee N, Shin B. Data quality management, data usage experience and acquisition intention of big data analytics[J]. International Journal of Information Management, 2014, 34(3):387-394.

[3] Hazen B T, Boone C A, Ezell J D, et al. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research

and applications[J]. International Journal of Production Economics, 2014, 154:72-80.

[4] Hu M, Lim E P, Sun A, et al. Measuring article quality in wikipedia: models and evaluation[C]//Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 2007:243-252.

[5] BLUMENSTOCK J E.Size matters:word count as a measure of quality on Wikipedia[C]//Proceedings of the 17th International Conference on World Wide Web．New York：ACM Press, 2008:1095-1096.

[6] Dalip D H, Gonçalves M A, Cristo M, et al. Automatic assessment of article quality in web collaborative digital libraries[J]. Journal of Data and Information Quality (JDIQ), 2011, 2(3): 14.

[7] Xu Y, Luo T. Measuring article quality in Wikipedia: Lexical clue model[C]//Web Society (SWS), 2011 3rd Symposium on. IEEE, 2011:141-146.

[8] Han J, Wang C, Jiang D. Probabilistic quality assessment based on article's revision history[C]//Database and Expert Systems Applications. Springer Berlin Heidelberg, 2011:574-588.

[9] YU S, MASATOSHI Y. Assessing quality scores of Wikipedia article using mutual evaluation of editors and texts[C]//Proceedings of the 22nd ACM Conference on Information and Knowledge Management．New York:ACM Press, 2013:1727-1732.

[10] Dalip D H, Gonçalves M A, Cristo M, et al. On multiview-based meta-learning for automatic quality assessment of wiki articles[M]//Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, 2012: 234-246.

[11] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[12] Verikas A，Gelzinis A, Bacauskiene M.Mining data with random forests：a survey and results of new tests[J]．Pattern Recognition,2011,44(2):330-349.

[13] LI Feng gang, LIANG Yu, GAO Xiaozhi, et al. Research on text categorization based on LDA-wSVM model[J].Application Research of Computers, 2015, 32(1): 21-25.

[14] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. the Journal of machine Learning research, 2003 (3):993-1022.

[15]Wikipedia: Version 1.0 Editorial Team/Assessment [EB/OL]. https://en.wikipedia.org/wiki/Wikipedia: Version _1.0 _Editorial_ Team, 2014.

[16] McCallum, Andrew Kachites. MALLET: A Machine Learning for Language Toolkit [EB/OL]. http://mallet.cs.umass.edu, 2002.