Thai Wikipedia Article Quality Filtering Algorithm

Nuanwan Soonthornphisaj and Peerapoom Paengporn

***Abstract*— Wikipedia is a content creation system that uses open collaboration as a strategy to drive the variety of topic coverage. There are approximately 100,000 Thai articles in Wikipedia. We found that the quality of content is the big issue since there are only 240 articles that has been labeled as featured articles whereas the rest of Thai articles are unlabeled. The website is ranked number 12 in term of user access in Thailand. That infers the use of their content in many academic documents and it would affect Thai educational quality in the long term. This paper present a good quality article filtering framework using decision tree algorithm. We propose new feature set obtained from a variety of references found in Wikipedia articles. The feature sets are applied in the machine learning algorithm in order to get the classifier with the knowledge concept of high and low quality articles. The performance of filtering algorithm on unlabeled articles is evaluated by real users to validate the performance of the system.**

***Index Terms*—Thai Wikipedia article, decision tree, feature set.**

I. Introduction

WIKIPEDEA web site contains highly dynamic information in term of articles. Users can share their knowledge by editing the content or creating new articles. The problem mostly found is the content quality problem since it is an open system that allow anonymous volunteers to edit any articles. The Wikipedia community has developed many policies and guidelines to improve the content quality of encyclopedia; however, it is found that many articles are not qualify. To guarantee the quality of the articles, the web site had launched the label so called *featured articles*. Featured articles are considered to be the best articles Wikipedia has to offer, as determined by Wikipedia's editors. They are used by editors as examples for writing other articles. Before being listed here, articles are reviewed as featured article candidates for accuracy, neutrality, completeness, and style according to the featured article criteria [1] Another label is called good articles which means that the article meets a core set of editorial standards but is not featured article quality. They are well written, contain factually accurate and verifiable information, are broad in coverage, neutral in point of view, stable, and illustrated, where possible, by relevant images with suitable copyright licenses. Good article does not have to be as comprehensive as featured article, but they should not omit any major facets of the topic [2].

Currently, there are 117 featured and 123 good articles written in Thai language. The quality of 97,452 articles are not yet defined by the community. We aim to use machine learning techniques to filter those unspecified articles that seem to be the candidate of good articles.

Our assumption is that the good writers should be honest and verifiable to the reader that means they should provide adequate reference sources for their articles. We propose the new feature set base on the concept of verifiable which is the common property of qualified articles that is useful for the learning algorithms to obtain the knowledge of good article and can be applied for the classification process. Since the unspecified quality articles has no label, therefore the predicted classes of the classifier are validated by real users as well.

The rest of the paper is organized as follows. In section II, state of the art is given. Section III deals with the filtering framework and feature selection method. The learning algorithms are described in Section IV. Experimental set up and results will be given in Section V. Conclusion drawn from the study have been given in Section VI.

II. Literature review

Text mining framework usually rely on the feature set that plays an important role to achieve the satisfactory performance. Considering the research work on Wikipedia data set, we found that feature set mentioned by researchers in literature can be summarized into three categories which are Review features, Network features, and Text features [3]. Review features are extracted from the review history of each article. An example of this feature is the Probability Review [5] which is used to assess the quality of a Wiki article based on the quality of its reviewers. Recursively, the quality of the reviewers is based on the quality of the articles they reviewed.

Network features are those extracted from the connectivity network inherent to the collection. An example

Manuscript received December 8, 2016; revised January 13, 2017. This work was supported in part by Kasetsart University Research Development Institute under Grant No. 125.59

Nuanwan Soonthornphisaj is an Associate Professor from Department of Computer Science, Faculty of Science Kasetsart University (corresponding author) phone: +662 562 5555; fax: +662942 8488; e-mail: fscinws@ku.ac.th).

Peerapoom Paengporn was the Computer Science student, Faculty of Science, Kasetsart University.

of this feature is out degree base on counting the number of links to other articles [6].

Text features are those extracted from the textual content of the articles. Examples of this feature are Structure features, Style features, and Readability features. Structure features are indicators of how well the article is organized. Style features are intended to capture the way the authors write the articles through their word usage. Readability features are intended to estimate the age or understandable grade level necessary to comprehend a text. The most frequent and second most frequent editor of the article are an example of this feature [7].

Based on the previous studies, several authors have proposed combining these features with Machine Learning method to represent the quality. For example, Naive Bayes Classifier, Decision Tree, k-means clustering algorithm, Support Vector Regression, and Support Vector Machines. For instance, [8] present a large number of features which are organized into three views of quality, related to the text of the article (e.g. its organization, length, readability), its revision history and network properties. These views are combined using a meta-learning strategy and Support Vector Regression. Factual density was used by [9] to measures the relative number of document facts and thus indicates a document's informativeness. They investigate the use of relational features for categorizing Wikipedia articles into featured/good versus non-featured ones base on a Naive Bayes Classier in combination with Information Gain feature selection. If articles have similar lengths, this methodology achieves an F-measure of 86.7% and 84% otherwise. Yanxiang and Tiejian [10] suggest a methodology for estimating the quality based on eight different ratios derived from counting the number of sentences, words, nouns, and other. They train a Decision Tree on a sample of 200 start class and 200 featured articles and test on a different sample of 100 start class and 100 featured articles, achieving precision and recall of more than 83% each. Liu and Ram [11] examine the quality of the articles to determine the impact of collaboration patterns on quality articles. This research applies K-means clustering algorithm and found that the collaboration of contributors' pattern is a critical factor driving the quality of Wikipedia articles. Dalip et al. [13] proposed a continuous quality scale based on a Support Vector Regression. Their observation is that the most useful feature is the text Structure. These features are easiest to extract. The best results are achieved when Structure features are combined with Network and Revision features. Lipka and Stein [12] present the character trigram feature, originally apply for writing style analysis. They combine a linear SVM with a trigram vector to achieve the performances in terms of the F-measure (0.964) for featured articles.

TABLE I
FEATURE SET

Feature	Meaning
# internal wiki links	Number of internal wiki links within the page
# external wiki links	Number of external wiki link to the current pages
# URL links	Number of external linked websites
# citations	Number of citations found in the content
# unique citations	Number of unique citation found in the content
Average number of citation per paragraph	Number of unique citation/number of paragraphs
Average size of content per reference	Number of character in the content/unique reference
External reference Popularity score	The average popularity score of external cited websites
Ratio between content size and # unique citation	Content size/number of unique citations
# paragraph	Number of paragraphs
# cited by wiki articles	Number of times that the article is cited by other Wikipedia articles.
# cited by redirected wiki article	Number of times that the article is cited by other redirected articles.
Content size	Number of characters
# paragraph	Number of paragraph

III. FILTERING FRAMEWORK

The framework for filtering the high quality articles has been set up as shown in Fig.1.

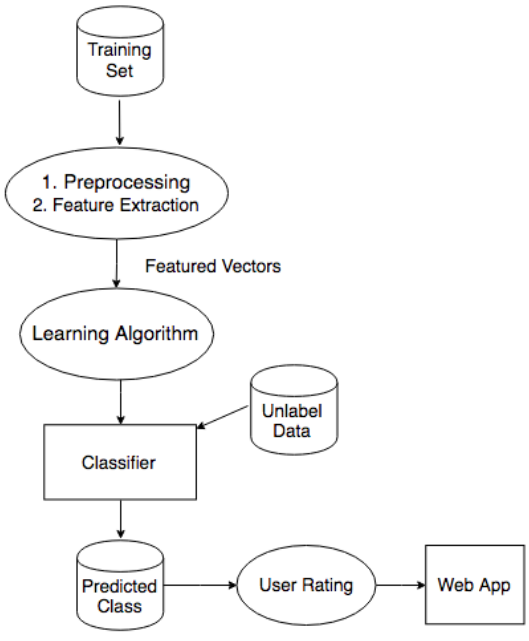

Fig. 1. Good Quality Article Filtering Framework

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Our assumption is that a good article should provide enough verifiable information and reasonable content size. Hence we propose a feature set based on various kind of references and the size of content as shown in Table I.

Given a Thai Wikipedia article, we extract three kinds of references; 1) internal wiki links within the page 2) inter wiki links that link to other pages and 3) Reference Websites. Another good practice in writing articles is to give credit to the original source of information using citation which infer how good is the writer and the quality of that article is inevitably consider as a good article. We extract the citation counts in term of unique citation and average of citation counts per paragraph.

Moreover the quality of external references is one of the most interesting feature. We develop an algorithm that calculate the popularity of each reference external website used by each domain, then for each Wikipedia article the average popularity score is obtained (see Table II)

TABLE II
FEATURE EXTRACTION ALGORITHM

Algorithm: PopularityScore (d: Thai Wikipedia Article)

Ref = {r_1,r_2,r_3,…r_n)
For each r_i in d
    RefScore = RefScore + GetPopularityURL(r_i)
    n++
PopularityScore = RefScore / n

Return PopularityScore

Algorithm: PopularityURL(D: dataset)

Create the list of unique URL found in dataset D
For each *d* in D
  If foundURL(r_i)
    Score ++
Return score

IV. LEARNING ALGORITHM

Decision tree is an algorithm that generates a tree representing the model of classes from training data. The algorithm is attractive because it can transform to the understandable set of rules. Each node in the tree is an attribute that is the best splitter because it can reduce the diversity of the predefined class in the training set by the greatest amount. The well-known decision tree proposed by Quinlan [3] namely C4.5 uses Gain ratio to avoid the bias caused by attribute having larger number of values.

$$Gain(S,A) = Entropy(s) - \sum_{v \in Values(A)} \frac{|S_v|}{S} Entropy(S_v) \quad (1)$$

Note that S is the prior data set before classified by attribute A, $|S_v|$ is the number of examples those value of attribute A are v, |S | is the total number of records in the dataset.

$$\text{GainRatio}(S, A) = \frac{Gain(S,A)}{SplitInfo(S,A)} \quad (2)$$

Where SplitInfo(S,A) is the information due to the split of S on the basis of the value of the categorical attribute *A*.

In this work, we apply J48 which is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool [14].

V. EXPERIMENT

A. Dataset

An XML dump file is obtained from Wikipedia Website. The preprocessing steps are as follows

1) The article extraction

The XML tag named <page>...</page>, <title>...</title> and <ns>0</ns> are used to identify the article. We get 221,100 Thai articles then we found that the set of articles contain redirected articles which means there is no content inside. Hence, we discard these redirected articles from our dataset using <redirect title="..." /> tag as a key to filter them out. Finally we got 97,752 articles for our experiment.

2) The class label determination

The ultimate goal of the encyclopedia website is of course to promote the quality of the content. Therefore the community has set up the rules to determine the quality. There are quite a number of tags that represents the quality of the articles such as clean up tag{{Cleanup}} which means that this article may require cleanup to meet Wikipedia's quality standards. Nevertheless most of articles have no tag that can be used to infer the quality of the articles. The class label determination step reveal the class distribution of the dataset in each domain as shown in Table III.

TABLE III
DATA SET

	Domain		
Label	Biography	Animal	Place
high quality	67	10	27
low quality	6462	244	4735
unlabeled	7627	892	5483

3) Feature extraction

Feature vector for each article in 3 domains are created. We explore the average of these feature sets to see the characteristics of articles in different labels and found that the average value of all features with the low quality label are lower than that of high quality whereas the mean of unlabelled class is in between the high and low quality class. Except for the Ratio between content size and # unique citation and External reference Popularity score.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE IV
AVERAGE VALUE OF FEATURE SET IN DIFFERENT CLASS LABELS

Domain	Biography			Animal			Place		
Feature	high quality	Low quality	Unlabeled	high quality	Low quality	Unlabeled	high quality	Low quality	Unlabeled
# internal wiki link	245.9	36.17	49.01	156.1	29.45	40.9	254.33	27.03	42.84
# cited by wiki articles	71.3	7.85	11.03	26.7	11.27	8.61	227.33	14.84	19.52
# URL link	57.52	3.02	6.25	22.8	2.24	4.06	70	2.79	5.43
# citation	95.91	2.35	6.62	67	2.51	6.15	92.52	1.56	3.77
# unique citation	7.81	2.06	6.62	40.5	2.02	4.83	75.33	1.27	3.17
Average number of citation per paragraph	1.55	0.35	0.54	1.02	0.44	0.63	1.04	0.31	0.37
External reference Popularity score	295.32	329.45	684.24	135.12	149.49	252.56	264.48	119.8	299.48
Ratio between content size and # unique citation	4465.18	5147.85	4464.33	2602.06	3065.36	2187.9	2003.57	5442.71	6079.51
Content size	122706.03	9345.76	13762.1	79101.4	6589.52	9705.4	115723.96	8616.07	14372.4

B. Evaluation

We are interested in precision and recall of the filtering algorithm. To evaluate these performances we use ground truth which obtained from Wikipedia community. Feature article and good article are combined to high quality label. The low quality label is also obtain based on community evaluation. Note that the unlabeled articles are used as a supplied test set and are manually evaluation by real user rating.

$$\text{Precision} = \frac{\#\text{ correctly retrieved articles}}{\#\text{ retrieved articles}} \quad (3)$$

$$\text{Recall} = \frac{\#\text{ correctly retrieved articles}}{\#\text{ relevant articles}} \quad (4)$$

TABLE V
THE PERFORMANCE OF FILTERING ALGORITHM ON PREDEFINED CLASS

	Biography	
	Precision	Recall
high quality	0.881	0.881
low quality	0.999	0.999
	Animal	
	Precision	Recall
high quality	0.909	1.00
low quality	1.00	0.996
	Place	
	Precision	Recall
high quality	0.857	0.889
low quality	0.999	0.999

TABLE VI
PERFORMANCE OF FILTERING ALGORITHM ON UNLABELED DATA

Predicted class	Biography	Animal	Place
high quality	100	24	55
low quality	7527	868	5428

TABLE VII
REAL USER RATING ON UNLABELED CLASS

Rating Score	3-3-3	3-3-2	3-2-2	2-2-2	Lower rating
# articles	105	43	20	5	6
percentage (%)	0.5866	0.2402	0.1117	0.0279	0.0335

Table V shows the performance of filtering algorithm for the predefined class. Note that, the dataset has two predefined classes which are high quality articles and low quality articles. (we combine two labels; feature article and good article as high quality class). The filtering algorithm is evaluate on 3 article domains which are Biography, Animal and Place. The precision obtained from high quality class of Biography, Animal and Place are 0.881, 0.909 and 0.857 respectively. The recall obtained from high quality class of Biography, Animal and Place are 0.881, 1.00 and 0.889 respectively. The decision trees obtained from the algorithm are shown in Fig.2, 3 and 4.

We do the second experiment by using unlabeled articles as a supplied test set. The filtering result is shown in Table VI. Since there is no predefined class in

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

unlabeled test set so we set up the real user rating system using 3 scales (3 = good quality, 2 = moderate, 1=low quality) in order to evaluate the performance of the filtering algorithm as shown in Table VII. There are 108 real users participate in this step to evaluate 179 articles which are filtered as good quality are validated to see the true positive rate of the filtering algorithm. Each article is evaluated by 3 users and we assume that if the article gets the good quality rating from at least 2 users it means that the article is really the good one. The result shows that 148 from 179 articles is really the good quality (the true positive is 82.68%)

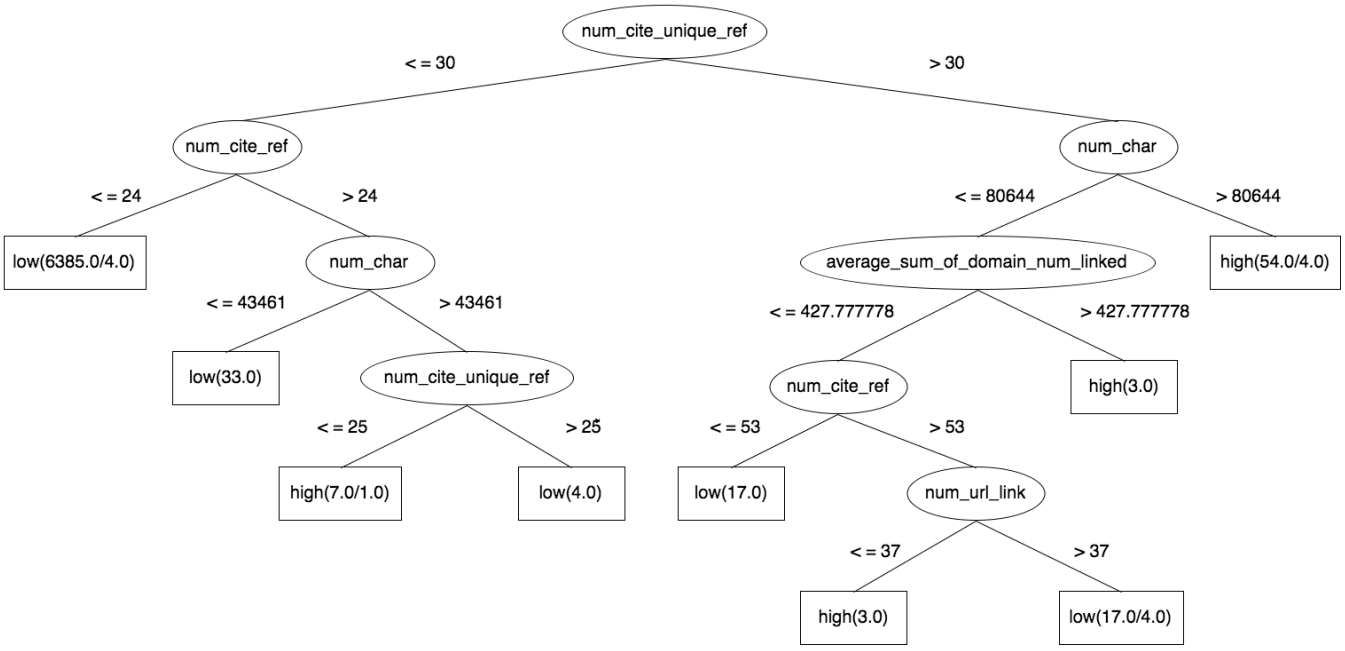

Fig. 2. Decision tree obtained from Biography domain

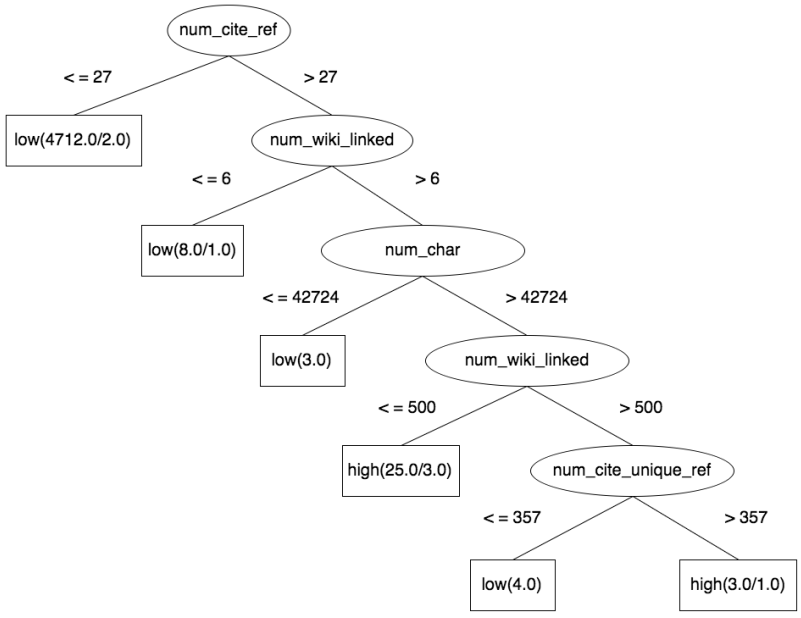

Fig. 3. Decision tree obtained from Place domain

Fig. 4. Decision tree obtained from Animal domain

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The performance in term of false negative is estimated from the low quality predicted by the algorithm. There are 13,828 articles that are classified as low quality. Since the number of these predicted article is quite high. Therefore we select the articles that has the tendency to misclassify as the low quality to be evaluated by real user rating. So we consider the feature set found in the decision tree as the key since these features have high value of information gain that infer the high classification power. These features obtained from decision tree are as follows

1) *# internal wiki links*
2) *# URL links*
3) *# unique citations*
4) *# external wiki links*
5) *External reference Popularity score*
6) *Content size*

We found that there are 525 articles that are classified as low quality but their feature values are higher than the average value of the predefined low quality class as shown in Table IV.

TABLE VIII
REAL USER RATING ON LOW QUALITY ARTICLES

Rating Score	3-3-3	3-3-2	3-2-2	2-2-2	Lower rating
# articles	23	46	74	48	334
percentage (%)	4.38	8.76	14.1	9.14	63.62

The real user rating shows that 13.14% of the predicted low quality class are contradicted with user opinion. We assume that if the article get good quality rating from at least 2 users it means the confirmation of the good quality articles. Therefore the false negative is 13.14%

VI. WEB APPLICATION

We have implemented the web application that facilitate user with 2 functions. 1) The keyword search function and 2) the user rating (Fig5, 6). After the search article is retrieved the quality of that article which obtained from the algorithm is shown. User can feedback the system with the quality rating so that the system can update the quality in term of statistics and can be used later in the learning process of the filtering framework.

Fig. 5 Main page of the filtering system

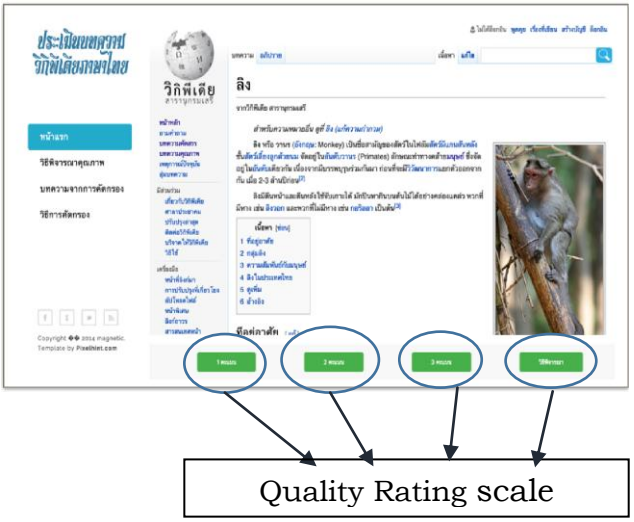

Fig. 6 The user rating scale

VII CONCLUSION

We found that feature sets based on references and citations show promising result in this work. The Article filtering system is implemented as a Web application. User can do keyword search to retrieve the Wikipedia articles. After browsing the article, user can give feedback in term of quality rating. We plan to combine the statistical feature with ontology to determine the article's quality in the near future.

ACKNOWLEDGMENT

This research is supported by Kasetsart University Research and Development Institute (KURDI), Grant No. 125.59, along with the Graduate school Kasetsart University and Department of Computer Science, Faculty of Science.

REFERENCES

[1] (2016) Wikipedia:Featured articles. [Online]. Available: http:// https://en.wikipedia.org/wiki/Wikipedia:Featured_articles
[2] (2016) Wikipedia:Good articles . [Online] . Available: https://en.wikipedia.org/wiki/Wikipedia:Good_articles
[3] D.H. Dalip, Gonçalves, M.A., Cristo, M., Calado P., 2011,"Automatic Assessment of Document Quality in Web Collaborative Digital Libraries," Journal of Data and Information Quality 2, pp.1-30.
[4] J. R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
[5] M. Hu, E-P.Lim, A. Sun, Lauw, HW. B. O.Vuong, "Measuring article quality in Wikipedia: Models and evaluation," In Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07). 2007, pp.243–252.
[6] L. Rassbach, T. Pincock, B. Mingus, "Exploring the Feasibility of Automatically Rating Online Article Quality," In Proceedings of the 9th Joint Conference on Digital Libraries.
[7] K. Saengthongpattana, N.Soonthornphisaj, "Thai Wikipedia Quality Measurement using Fuzzy Logic," In 26th Annual Conference of the Japanese Society for Artificial Intelligence, June 12–15, 2012, Yamaguchi, Japan.

[8] D. H. Dalip, M.A. Gonçalves, T. Cardoso, M. Cristo, P. Calado,"A Multi-view Approach for the Quality Assessment of Wiki Articles," Journal of Information and Data Management, 2012, vol. 3 no.1, 10.
[9] E. Lex, M.Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein and M. Granitzer, "Measuring the quality of web content using factual information," In Proc. of WICOW. 2012, pp.7–10.
[10] X. Yanxiang, and L. Tiejian, "Measuring article quality in Wikipedia: Lexical clue model," 3rd Symposium on Web Society, 2011 IEEE, pp.141–146.
[11] J. Liu,, and S.Ram, "Who does what: Collaboration patterns in the Wikipedia and their impact on article quality," ACM Trans. Manage. Inf. Syst, vol. 2, 2011, pp.1-23.
[12] N. Lipka, Stein, B., "Identifying featured articles in wikipedia: writing style matters," Proceedings of the 19th international conference on World wide web, ACM, Raleigh, North Carolina, USA, 2010, pp.1147-1148.
[13] D.H. Dalip,, Gonçalves, M.A., Cristo, M., Calado P.,"Automatic Assessment of Document Quality in Web Collaborative Digital Libraries," Journal of Data and Information Quality, vol.2, 2011, pp.1-30.
[14] E. Frank, M. A. Hall, and I. H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

ISBN: 978-988-14047-3-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)