Automation, Control and Intelligent Systems
2015; 3(6): 141-153
Published online January 15, 2016 (http://www.sciencepublishinggroup.com/j/acis)
doi: 10.11648/j.acis.20150306.18
ISSN: 2328-5583 (Print); ISSN: 2328-5591 (Online)

Qualifying Articles of Persian Wikipedia Encyclopedia Through J48 Algorithm, ANFIS and Subtractive Clustering

Seyedtaha Seyedsadr[1], Mohammadali Afsharkazemi[2], Hashem Nikoomaram[3]

[1]Department of Management, Electronic Branch, Islamic Azad University, Tehran, Iran
[2]Department of Management, Tehran Central Branch, Islamic Azad University, Tehran, Iran
[3]Department of Management and Economics, Sciences and Research Branch, Islamic Azad University, Tehran, Iran

Email address:

Sts.sadr@srbiau.ac.ir (S. Seyedsadr), m.ali.akazemi@gmail.com (M. Afsharkazemi), nikoomaram.hashem@gmail.com (H. Nikoomaram)

To cite this article:

Seyedtaha Seyedsadr, Mohammadali Afsharkazemi, Hashem Nikoomaram. Qualifying Articles of Persian Wikipedia Encyclopedia Through J48 Algorithm, ANFIS and Subtractive Clustering. *Automation, Control and Intelligent Systems.* Vol. 3, No. 6, 2015, pp. 141-153.
doi: 10.11648/j.acis.20150306.18

Abstract: Since Wikipedia encyclopedia is one of the most popular web sites on the internet, providing accurate information is of abundant importance. In this research, the effective variables on quality of Persian articles are identified and a system is, then, designed for judging articles in three quality levels: high quality, cleanup needed, and deletion. First, the variables relating to the articles included in the list of featured articles, good articles, cleanup needed, and deletion articles are collected. Then, two methods are used for the analysis of data: First, a decision tree explains the relationships among the collected variables as rules that are implemented by adaptive neuro fuzzy interference system. Second, the data are implemented by subtractive clustering algorithm and the error of both methods is, finally, measured and compared. The results indicate that the average daily hits, total views, page length, total number of edits, total number of authors, and number of templates used are directly related to quality of Persian articles while the number of recent number of authors is inversely related to quality of articles.

Keywords: Wikipedia Encyclopedia, Quality of Articles, J48 Decision Tree, ANFIS, Subtractive Clustering Algorithm

1. Introduction

Persian Wikipedia is a web-based free online encyclopedia that was founded in December 2003 [1]. This encyclopedia included 373512 articles, 330083 users, and 28 administrators in October 31, 2013 which won the 20th place among the various encyclopedia languages [2]. Free access, constant updates, extensive coverage, and diversity have turned Wikipedia into one of the most unmatched social websites in the cyberspace [3]. Increasing the number of members of a group may usually reduce the effectiveness of teamwork; however, the researchers have assessed wisdom of crowds and collective action as the success factors of Wikipedia [4, 5, 6]. The users residing in the Europe and North America have mostly contributed in writing the articles of Persian Wikipedia and the Iranian users' contribution is only about 45% [1]. Most people who have contributed in the development of Wikipedia are anonymous [1, 7]. Users with different motives have helped development of this encyclopedia. The new users help development of this encyclopedia to satisfy their curiosity and the old users do so with such motivations as providing information and producing content. Sense of usefulness, sense of finding identity in the society, feedbacks received from the user community, and achieving fame are considered among the most important factors of continuation of contribution of the Internet users with the development of the Persian Wikipedia. Financial incentives have not had any role in collaboration of the users with the Persian Wikipedia [1]. Internet users through having access to the pages of Wikipedia without any limitation and make some changes in those pages except in cases where a particular page is restricted to be edited by the ordinary users by the site administrators [8]. Many people use the site information every day, develop new articles, and edit the older articles. Hence, one of the most important challenges of the site is to deliver an optimum quality level of the articles. People with different ages, cultures, viewpoints, and opinions can express their views on different issues in this encyclopedia regardless of what their specialty is; thus, Wikipedia has prescribed instructions and guidelines for improving the quality of this encyclopedia [9, 10]. Generally, the quality of articles in the site of Wikipedia is determined using opinion polls [11]. Further, the associated

research on determining the quality of articles of Wikipedia encyclopedia and using decision tree, ANFIS, and subtractive clustering algorithm in the predictions are examined.

2. Related Work

2.1. Determining the Quality of Wikipedia Articles

Walraven et al. [12] would demonstrate that students and university students spend most of their time searching for information in the cyberspace while they spent less time for assessment and evaluation of the contents. Conducting a research on method of study of a number of students of an American college, Lim [13] suggests that students often search the basic information of the intended subject in the Wikipedia and this encyclopedia directs them towards the use of more specialized sources, in this study, more effort is recommended to improve the quality, precision, and accuracy of the Wikipedia information as well. Lucassen and Schraagen [14] described in a research carried out on the views of the 149 people from different parts of the world that the users do not highly trust on the Wikipedia articles. Wikipedia has been mainly criticized for the low information quality. Stvilia et al. [15] have assessed the information quality criteria in the English Wikipedia by studying the manner of selection of the featured articles and demonstrated that information quality could not be measured with a simple model. Geiger and Halfaker [16] examined the impact of the Wikipedia software robots to prevent vandalism and control the quality of this encyclopedia. Rowley and Johnson [17] have demonstrated that multiplicity of authorship, references, proper structure and grammar would increase the trust of addressees as users while the comments representing the concerns such as citation needed have a negative impact on their trust. Noč and Zumer [18] examined the general situation and the quality of featured articles of the Slovenia Wikipedia based on reviewing the references and sources and concluded that the quality of the featured articles were generally higher than other articles. Kittur and Kraut [4] have suggested a method for improving the quality of articles through coordinating the editors and substantiated the articles that had more editors were of higher quality than the other articles. Improving quality of the articles is achieved, when the editors use coordination techniques. Ram and Liu[3] have examined the effect of the variables of collaboration patterns, number of unique editors, number of edits, article age, article length, and number of unique administrators on quality of articles and indicated that the articles written by professional authors have often higher quality than other articles. Lih [19] measured the quality of Wikipedia articles by the independent variables of linking, number of edits, number of authors, article size, and other metadata from Wikipedia. Priedhorsky et al. [5] have examined the impact of editing and editors on the Wikipedia value. Stein and Hess [20] have demonstrated that there is a significant relationship between the featured articles and increasing number of authors. In a research conducted on the featured articles of the English Wikipedia, Wilkinson and Huberman [6] have substantiated that there is a direct relationship between increasing the number of editions, the number of authors, contribution in editing the articles, and higher quality. Wöhner and Peters [21] have separated featured and good articles (high quality) from deletion articles (low quality) through examination of the two variables of transient contribution and persistent contribution and calculation of the lifecycle of articles. Saengthongpattana and Soonthornphisaj [22] have developed a system for separating the featured articles from the ordinary ones using fuzzy logic and K-mean algorithm through study of specifications of the featured articles in the Thai Wikipedia such as the number of images, links, main headings, footnotes, etc. Most of the researches conducted on the encyclopedia articles quality have focused on identifying and selection of the featured articles, Anderka et al.[23] have examined the cleanup needed articles and the tags for improving quality of the encyclopedia articles. Xiao et al.[24] separated the featured articles from the start class by C4.5 algorithm using statistical properties and data of the articles such as page length, number of authors, number of editors, number of links, and number of images. Chai et al. [25] presented a model for the assessment and measuring the quality of the posts sent by users at the low, medium and high levels.

2.2. Using the J48 Decision Tree and Fuzzy Algorithm in Predictions

Saravanan et al. [26] have presented a method for determining the status of inaccessible gears in a device. They processed the vibrating signals extracted from the gears and via selecting, the best statistical features of audio signals by J48 decision tree and then implemented the resulting rules using fuzzy logic toolbox of MATLAB software. Therefore, the developed system could detect and distinguish to the defected gears from the sound gears by the audio signal. Omid [27] developed a system using the J48 decision tree and fuzzy algorithm that recognizes the natural open pistachio from the closed pistachio. Jalili and Mahmoudi [28] have developed a system using the J48 decision tree and fuzzy algorithm that separates the two types of Iranian pistachios-Akbari and Kaleghouchi. Noorallah et al.[7] have judged the quality of 226 German Wikipedia articles in a research and assessed low quality articles with (0) and high quality ones with (1). Then, the statistical properties of the articles such as page length, number of authors, number of editors, etc., are classified by J48 decision tree of Weka software. The resulting rules are implemented by adaptive neuro-fuzzy inference system; thus, the researcher develops a system that separates high quality articles from the low quality articles like the human brain.

2.3. Subtractive Clustering Method

Clustering data of a collection aims at a brief display of the behavior of the dominant system in that collection. A set of input and output data of the system are collected. Indeed, the cluster centers specify the desired system behavior. Therefore,

each of the cluster centers can be regarded as a base of a rule which is used to describe the intended system behavior [29]. Fuzzy C-Mean clustering Method has numerous applications in unsupervised classification. One of the problems of this method is estimation of the initial quantity of cluster centers [30]. Like many non-linear optimization problems, quality of the FCM solution depends on the initial number and quantity of the cluster centers [29]. Yager and Filev and Yager [30] in 1992 developed a quick method for approximate clustering and estimation of the cluster centers by mountain method. Chiu [29] in 1994 examined and modelled the relationship between five independent variables of population, number of residential units, number of automobile owners, average household income, number of servants and the dependent variable of number of trips by subtractive clustering algorithm in Delaware, New Castle County [29]. A six-dimension space with three rules was covered by SC method and Chiu indicated that SC method reduced computational complexity to a great extent [29, 31]. The greater the radius of clusters, the less the number and centers of clusters will be using SC algorithm; thereby the number of rules will be reduced. The less the radius of clusters, the greater the number of clusters and so the greater the number of rules will be achieved. One of the advantages of SC method is that it is not needed to estimate the clusters [32].Yuan et al. [33] have predicted the quality of software in a research via using SC method. Wei et al.[34] have developed a model for prediction of the stock market of Taiwan by SC method. Malhotra and Sharma [35] have examined and modelled the relationship between nine independent variables such as the number of words in web pages, page size, the number of tables, graphs in the page, etc. and the dependent variable of quality of web pages using SC method. In another research, Afshoon et al. [36] have examined the quality of educational websites by SC method.

2.4. Summing Up

Considering the increasing number of contents produced by the users in the cyberspace, it is difficult to access high quality content. Determining the quality of the produced content is an inevitable requirement [25]. The earlier studies have revealed the necessity of conducting a research on quality of the Wikipedia articles in all languages including Persian. Most of the researches conducted on quality of the Wikipedia articles have assessed the articles of this encyclopedia in two high and low quality levels e. g. [3, 4, 6, 7, 19, 20, 21, 24]. cleanup needed articles cover a part of the Wikipedia articles in which some of the encyclopedia criteria are not observed [37]. These articles cover the intended subject reasonably; however, more references may be needed or the edition instruction of the Wikipedia may not be completely adhered to. As suggested by Anderka et al., this group of articles is less analyzed by the researchers [23, 37]. This research aims at answering to this question, “Do the seven variables (1) average visits of an article per day, (2) total number of visits of an article (view), (3) page length, (4) total number of edits, (5) total number of distinct authors, (6) recent number of distinct authors (rec authors), and (7) number of transcluded templates in an article affect the quality?”, it also aims at designing and implementation of a system that can judge the Persian Wikipedia articles in three quality levels of high, cleanup, and deletion like users and administrators. Figure 1 shows the effective variables on quality of articles of the Persian Wikipedia. Considering the earlier research and potentials of the Persian Wikipedia, the variables of the available articles are collected and saved in the list of featured articles, good articles, cleanup needed articles and articles for deletion. Two methods are used to analyze the obtained data. In the first method, the collected data are classified using the J48 decision tree of Weka software and the relationships between the dependent and independent variables are extracted in the form of rules with if-then structure. The resulting rules are, then implemented by adaptive neuro fuzzy inference system (ANFIS). The designed system separates the Persian articles of the Wikipedia in three high quality, cleanup, and deletion levels, and finally the designed system error is measured. In the second method, the variables data are processed, clustered, and implemented without using the decision tree by subtractive clustering algorithm. The designed system error is measured and compared with the first method. Figure 2 shows the framework of method 1 and 2. Further, the dependent and independent variables and the tools applied in the research are introduced. In section 4, the J48 decision tree is used for processing data and ANFIS and SC algorithm are used to implement the model. The obtained results are analyzed in section 5.

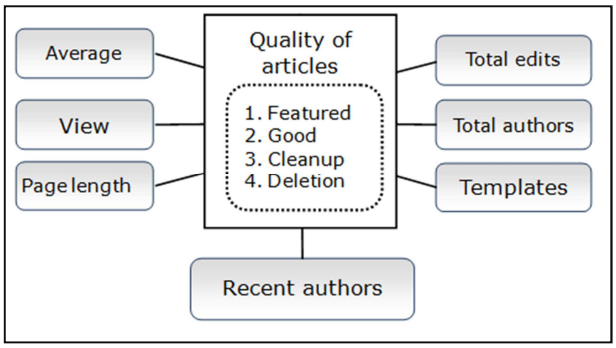
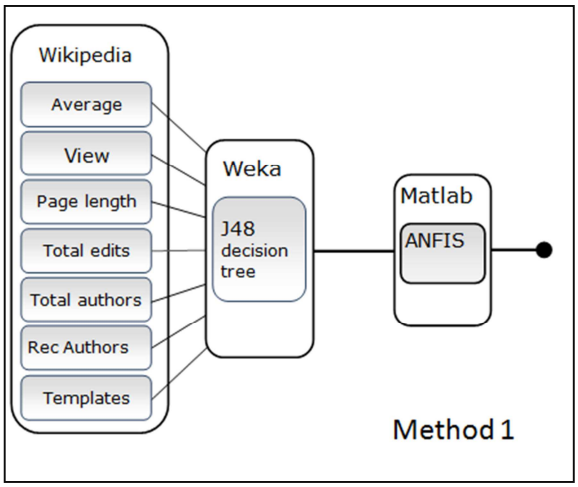


Figure 1. *Research variables model.*



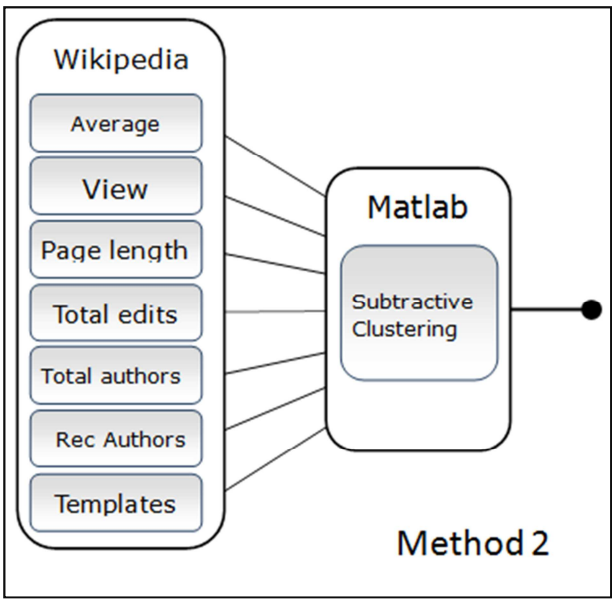


Figure 2. *Methods framework.*

3. Variables and Tools

The research data are the information related to the articles in the Persian Wikipedia that were extracted from the database of the Wikipedia site in October and November 2013. The purposive sampling method has been used in this research [38].

3.1. Dependent Variables

3.1.1. Articles for Deletion

The Wikipedia site deletes the articles which do not meet the criteria of the encyclopedia contents. Generally, the articles that ignore the copyright or cannot be attributed to reliable sources, new words, violation of biography of those who are alive or advertising are included in the cases of deletion. A feedback form is mainly used to remove articles from the Wikipedia site and users can discuss regarding deletion of articles. The discussion will last at least 7 days. Then, the article will be deleted if the users are agreed to do so [39]. In this research, 75 articles have been examined under the title of deletion articles variable.

3.1.2. Cleanup Needed Articles

Cleanup includes correcting spelling, grammatical, and typographical mistakes, inappropriate tone, and failure to use proper references. Contribution of all users is welcomed in edition or amendment of these articles. Finally, the changes made by a user on an article are discussed and assessed by other users [37]. In the present research, 52 articles in this list have been reviewed under the title of cleanup articles variable.

3.1.3. Good Articles

Good articles of the Persian Wikipedia are the articles with a good prose that are provable, have reliable sources, appropriate coverage of the content, unbiased attitude, and appropriate and relevant image. These articles are first nominated for the opinion poll of other users and are, then, selected as a good article in case of agreement of users. good articles are the complete articles though they have some problems, too [40, 41]. Sixty-eight good articles have been reviewed in the present research under the title of good articles variable.

3.1.4. Featured Articles

Featured articles are the professional, unique, and comprehensive articles that can be considered as the final source for information of the Wikipedia [42]. The featured articles shall be already selected as good comprehensive articles, with the volume of more than 5 KB; they shall have valid sources and must avoid addressing the marginal unnecessary issues, the Wikipedia instruction must be adhered to in these articles and they shall not be needed to be corrected. The articles are first nominated for the opinion poll of users and are selected as the featured article in case of agreement of users [43]. 89 featured articles under the title of featured articles variable have been analyzed in this research.

3.2. Independent Variables

(1) Average.

A numerical quantification obtained from division of the number of times users visit a particular article into the number of elapsed days since the release of that paper (average=view/days). Specifying the value of this variable for each article is possible using the Toolserver tools [44, 45].

(2) Views.

A numerical quantification, which displays the number of times users visit a specific article. Specifying the value of this variable for each article is possible using the Toolserver tools [44, 45].

(3) Page length.

The space occupied by the article on the encyclopaedia server, and its unit of measurement is bytes.

(4) Total number of edits.

The number of times that the article is written and edited [46].

(5) Total number of authors.

The number of people who have edited the article [46].

(6)Recent number of distinct authors (within the last 30 days).

The number of users who have written or edited the article within the last thirty days or month [46].

(7) Transcluded templates.

It is referred to the codes saved in the template namespace and is used to display the tables at the top and bottom of the article or the flags, etc. templates have also served as function in the programming languages [46].

3.3. Tools

3.3.1. Decision Tree

Decision tree is a unique method of presenting a system, which defines the intended system properly and simplifies the associated decisions [26, 27].

3.3.2. Fuzzy Inference Systems

Fuzzy inference is a process that converts the input into output by the defined rules. The basic structure of the fuzzy inference systems is composed of three sections. section one includes rules that is in the if-then form; section two is a

database through which the Membership Functions (MFs) are defined, and finally section three includes the inference mechanism that is achieved with the help of the rules and use of the available data to attain a reasonable output.

Fuzzy inference systems have displayed a successful performance in the areas of data classification, decision analysis, and expert systems [47].

- Takagi-Sugeno Inference systems model.

One rule has been shown in the Sugeno fuzzy model in Eq. 1.

(Eq.1) If x is A and y is B then z=f(x, y)

A and B are fuzzy sets, f(x, y) is a function in the conclusion section of the rule. When f(x, y) is a first-order polynomial, the resulting fuzzy inference system is called first-order Sugeno fuzzy model. In addition, if f (x, y) is equal to a fixed amount, the resulting fuzzy inference system is called zero-order Sugeno fuzzy model [48].

- ANFIS.

ANFIS is an acronym made from the first letters of Adaptive Neuro Fuzzy Inference System [31]. Qualitative aspects of human knowledge can be modelled by ANFIS [34]. ANFIS is a hybrid system that takes advantage of the potential benefits of Fuzzy Inference System (FIS) and Artificial Neural Network. Dynamic systems do not usually have linear behavior. Dynamic systems can be modelled with a linear behavior using. Sugeno fuzzy systems can be developed through ANFIS toolbox and tested after training [32].

- Subtractive Clustering method.

Clustering aims at categorizing a very large data set and providing a simple representation of the system behavior [31]. In case no information is available regarding the number of data set clusters, using SC algorithm is a quick. method for estimating the number of clusters and determining their centers [31]. The clustered data can be used to create a fuzzy inference system with the minimum rules and maximum efficiency [29]. ANFIS toolbox supports SC method [31]. In the next part, the variables are processed, classified, and clustered using version 3.6.8 of Weka and version 2, 1, 1 of fuzzy toolbox of MATLAB 8.3.

4. Methods and Results

4.1. Integration of Featured and Good Articles

First, the variables relating to 89 featured articles and 68 good articles are entered Weka in a CSV file format and are classified by the J48 decision tree. Figure 3 indicates that the statistical properties of the featured and good articles are so similar that cannot be separated. Given the similarity of specifications of the good and featured articles of the Persian Wikipedia, both of these articles have been examined under the title of a new dependent variable called high quality in this research.

Given the similarity of specifications of the good and featured articles of the Persian Wikipedia, both of these articles have been examined under the title of a new dependent variable called high quality in this research.

```
=== Summary ===

Correctly Classified Instances          89               56.6879 %
Incorrectly Classified Instances        68               43.3121 %
Kappa statistic                          0
Mean absolute error                      0.4911
Root mean squared error                  0.4956
Relative absolute error                 99.9748 %
Root relative squared error            100.0001 %
Total Number of Instances              157

=== Confusion Matrix ===

  a  b   <-- classified as
 89  0 |  a = Featured
 68  0 |  b = good
```

Figure 3. *Validation and confusion matrix results for good and featured articles.*

```
Correctly Classified Instances         254               89.4366 %
Incorrectly Classified Instances        30               10.5634 %
Kappa statistic                          0.8203
Mean absolute error                      0.0746
Root mean squared error                  0.2469
Relative absolute error                 18.889  %
Root relative squared error             55.6181 %
Total Number of Instances              284

= Confusion Matrix ===

  a   b   c   <-- classified as
 71   4   0 |   a = Deletion
  6  35  11 |   b = Cleanup
  0   9 148 |   c = High quality
```

Figure 4. *Validation and confusion matrix results for high quality, cleanup and deletion articles.*

4.2. Method 1

4.2.1. Classification with the Decision Tree

Variables average, view, page length, number of edits, number of authors, recent authors, and templates are entered Weka in a CSV file format and are classified by the J48 decision tree. Figure 5 shows the decision tree [49].

As seen in figure 4., accuracy=89.4% and kappa statistic = 0.82 that is an acceptable result [50].

Confusion Matrix is shown in figure 4. ab and aa elements of CM matrix reveal that 71 articles have been classified correctly and 4 articles have been classified incorrectly from among 75 deletion articles. ca and ac elements reveal that none of the deletion and high quality articles have been classified incorrectly. cc and cb elements indicate that 148 articles have been classified correctly and 9 articles have been classified incorrectly from among 157 high quality articles. bc, bb, and ba indicate that 35 articles have been classified correctly and 17 articles have been classified incorrectly from among 52 cleanup articles. Given the results of CM matrix, it is concluded that deletion and high quality articles have been separated without any error [50]. The results obtained from J48 decision tree are as follows, see figure 5:

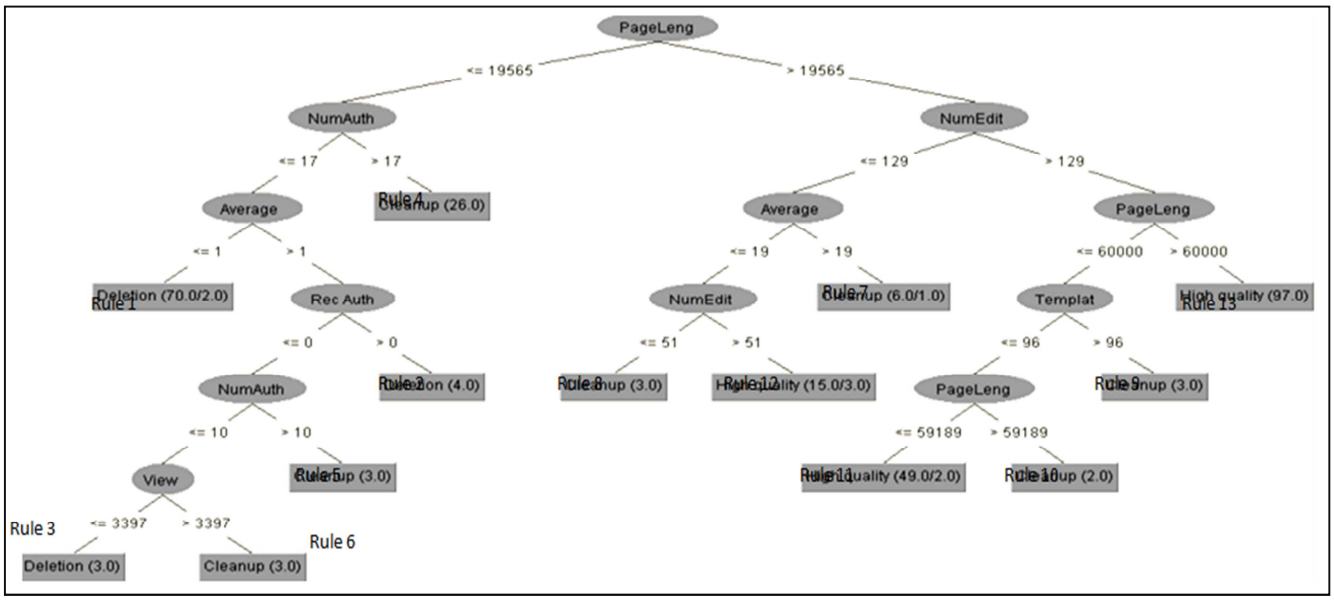


Figure 5. *Decision tree for J48 algorithm.*

4.2.2. Rules

Rule 1.

If (Average=<1) and (NumAuth=<17) and (PageLeng=<19565) then article is Deletion.

Rule 2.

If (Average>1) and (NumAuth=<17) and (PageLeng=<19565) and (RecAuth>0) then article is Deletion.

Rule 3.

If (Average>1) and (View=<3397) and (NumAuth=<10) and (PageLeng=<19565) and (RecAuth=<0) then article is Deletion.

Rule 4.

If (PageLeng=<19565) and (NumAuth>17) then article is Cleanup.

Rule5.

If (Average>1) and (PageLeng=<19565) and (NumAuth>10) and (RecAuth=<0) then article is Cleanup.

Rule 6.

If (Average>1) and (View>3397) and (PageLeng=<19565) and (NumAuth=<10) and (RecAuth=<0) then article is Cleanup.

Rule 7.

If (Average>19) and (PageLeng>19565) and (NumEdit=<129) then article is Cleanup.

Rule 8.

If (Average=<19) and (PageLeng>19565) and (NumEdit=<51) then article is Cleanup.

Rule 9.

If (PageLeng=<60000) and (NumEdit>129) and (Templat>96) then article is Cleanup.

Rule 10.

If (PageLeng>59189) and (NumEdit>129) and (Templat=<96) then article is Cleanup.

Rule 11.

If (PageLeng=<59189) and (NumEdit>129) and (Templat=<96) then article is High quality

Rule 12.

If (Average=<19) and (PageLeng>19565) and (NumEdit>51) then article is High quality.

Rule 13.

If (PageLeng>60000) and (NumEdit>129) then article is High quality.

4.2.3. Implementing with ANFIS

- ANFIS inputs.

The structure of Sugeno rules is the same as the rules obtained by the J48 decision tree. Seven variables are defined as ANFIS inputs [7, 26, 27]. Given the simplicity of relationships and calculations, input Membership Functions are taken into account as trapezoidal [26, 27]. Three MFs have been defined for variable average, two MFs have been defined for variable view, four MFs have been defined for variable Page length, three MFs have been defined for variable number of edits, three MFs have been defined for variable Number of Authors, two MFs have been defined for variable Recent Authors, and two MFs have been defined for variable templates.

- ANFIS output.

Thirteen MFs are allocated to the output, of which 3 MFs are allocated to deletion articles, 7 MFs are allocated to Cleanup articles, and 3 MFs are allocated to high quality articles. Each rule has been specified with a numeric value. The numeric values between -1 and +1 are used to specify the output MFs in a sense that the MFs with the values close to +1 indicate high quality articles, the MFs with the values close to 0 indicate Cleanup articles, and the MFs with the values close to -1 indicate deletion articles. Figure 6, shows the structure of designed system.

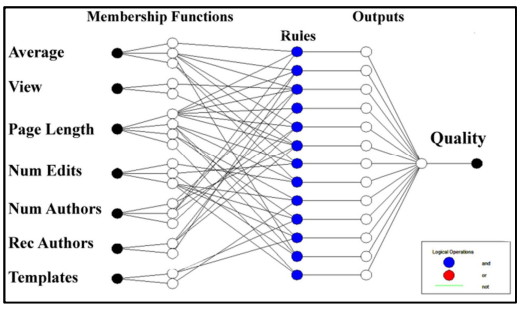


Figure 6. *System structure in method 1 (using ANFIS and J48 decision tree).*

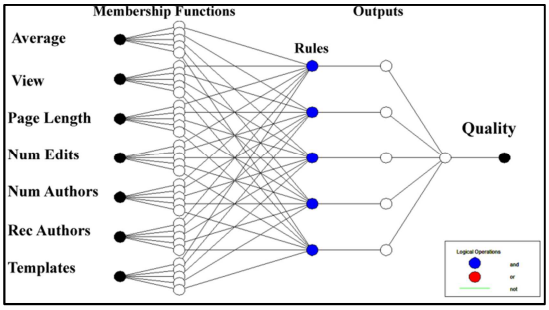


Figure 7. *System structure in method 2 (using SC algorithm).*

4.2.4. Model Validation

After designing the system, 189 articles (66% of data) have entered the system under the title of training data and 95 articles (33% of data) have entered under the title of checking data [31]. Figure 8, shows the output of the designed system when epoch = 0. The output of designed system with (○), and checking data with (-) are shown, respectively. Root Mean Square Error for training data and checking data are shown in (1) below.

(1) Epoch=0, Training RMSE=0.34, Checking RMSE=0.33.

As seen in figure 9, the error is decreased by increasing epoch. Considering figure 9, when Epochs>600, reduction of the error of checking data is so insignificant that has actually no remarkable impact on reduction of the system error; therefore, the system training is suspended in this stage and the RMS Error values of training and checking data are shown in (2) and the system output is shown in figure 10.

(2) Epochs=600, Training RMSE =0.271, Checking RMSE= 0.279.

Comparison of figures 8 and 10 shows that high quality and cleanup needed articles are more concentrated on the range of 0 and 1 while deletion articles are more scattered around -1. However, the system error is reduced as shown in (2).

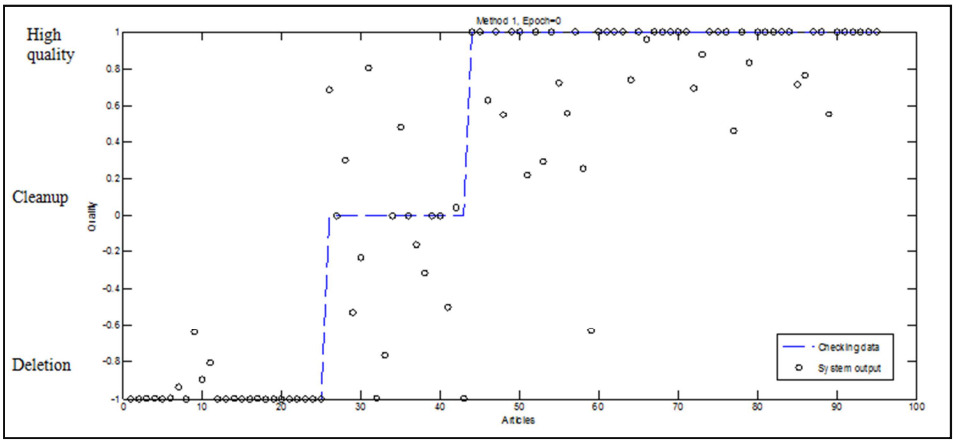


Figure 8. *The model output and checking data are shown as circles and solid blue line, when epoch=0, in method 1.*

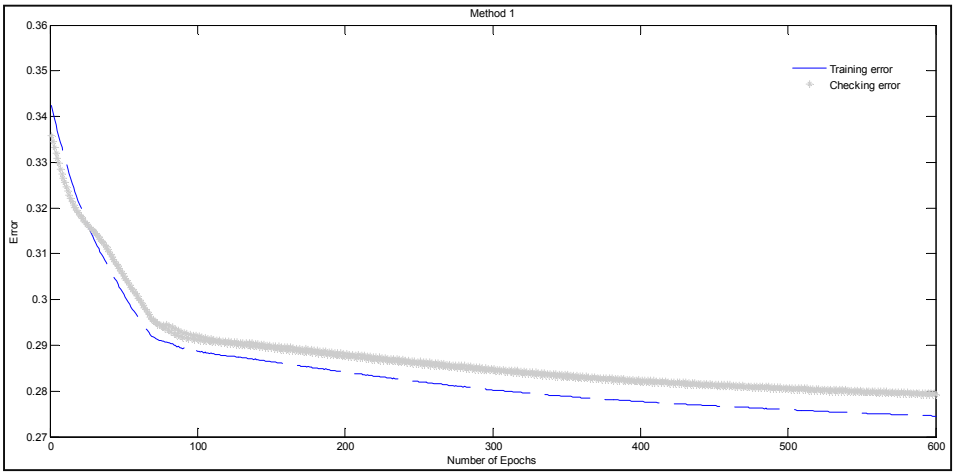


Figure 9. *RMS Error values of training and checking data.*

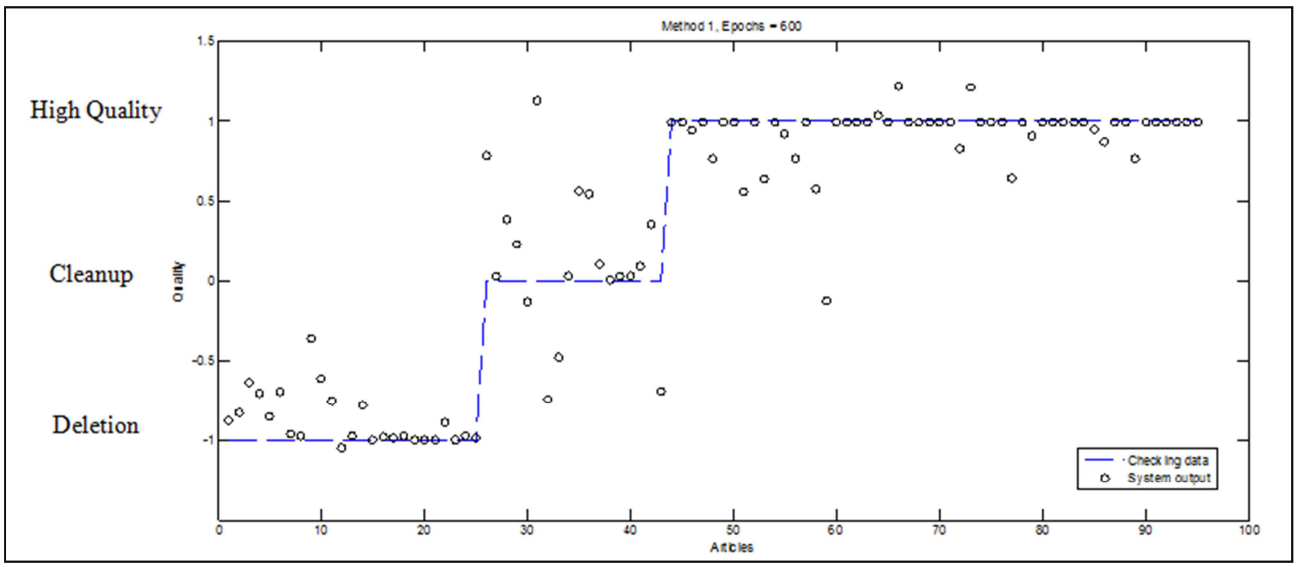


Figure 10. *The model output and checking data are shown as circles and solid blue line, when epoch=600, in method 1.*

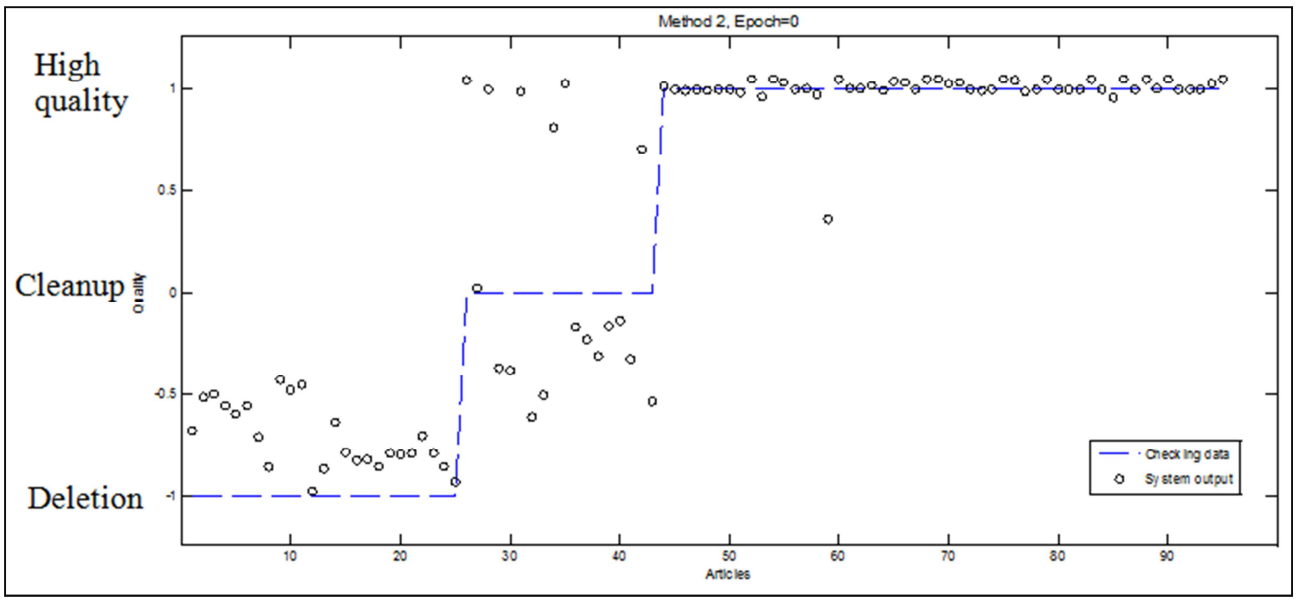


Figure 11. *The model output and checking data are shown as circles and solid blue line, when epoch=0, in method 2.*

4.3. Method 2

- Implementation with Subtractive Clustering

The membership function used in this method is Gaussian. The independent variables of average, view, page length, and number of edits, number of authors, recent authors, and templates enter the Workspace of MATLAB as Wikidatain and the quality of Wikipedia articles enter as Wikidataout. Instead of using of the titles of deletion, cleanup, and high quality articles, the numerical values of -1, 0, 1 are used, respectively.

(3) [C55, S55]=Subclust ([wikidatain, wikidataout], 0.55).

The command seen in (3) has allocated 5 MFs to each input variables and 5 MFs to the output variables, of which 1 MF is allocated to deletion articles, 1 MF is allocated to cleanup articles and 3 MFs are allocated to high quality articles. Value 0.55 indicates the radius of each cluster [31]. Figure 7, shows the designed system.

After designing the system, 189 articles (66% of data) have entered the system as training data and 95 articles (33% of data) have entered as check data. When epoch=0, RMS Error for training data and checking data are shown in (4). Figure 11, shows the output of the system, when epoch = 0.

(4) Epoch = 0, Training RMSE = 0.31, Checking RMSE= 0.32.

When epochs=600, RMSE for training data and checking data are shown in (5). Figure 12, shows the output of the system.

(5) Epoch = 600, Training RMSE = 0.27, Checking RMSE= 0.29.

Figure 13, indicates that the articles quality is improved by increasing average, view, page length, number of edits, number of authors, and templates variables. Figure 14, indicates that the variable recent authors is reduced by improving the articles quality.

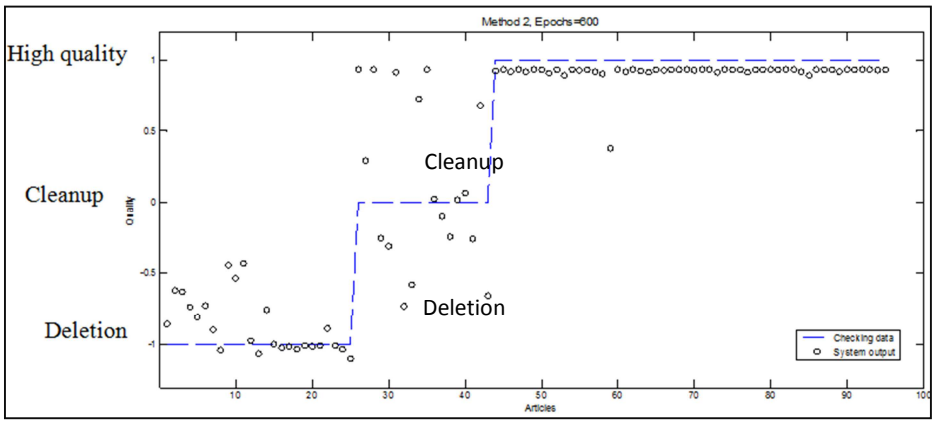


Figure 12. *The model output and checking data are shown as circles and solid blue line, when epoch=600, in method 2.*

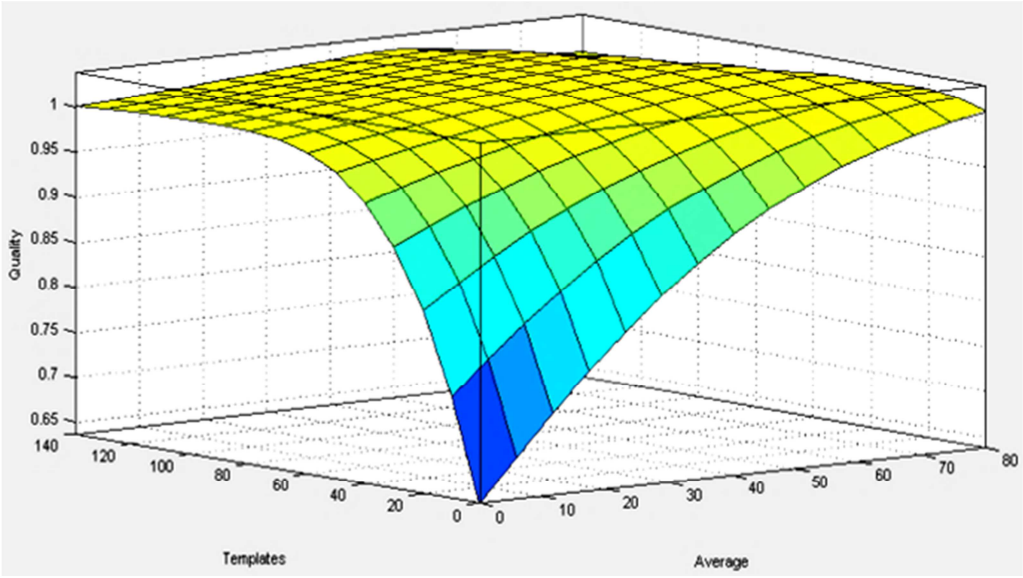


Figure 13A. *The increase in Average and Templates variables cause an increase in articles quality.*

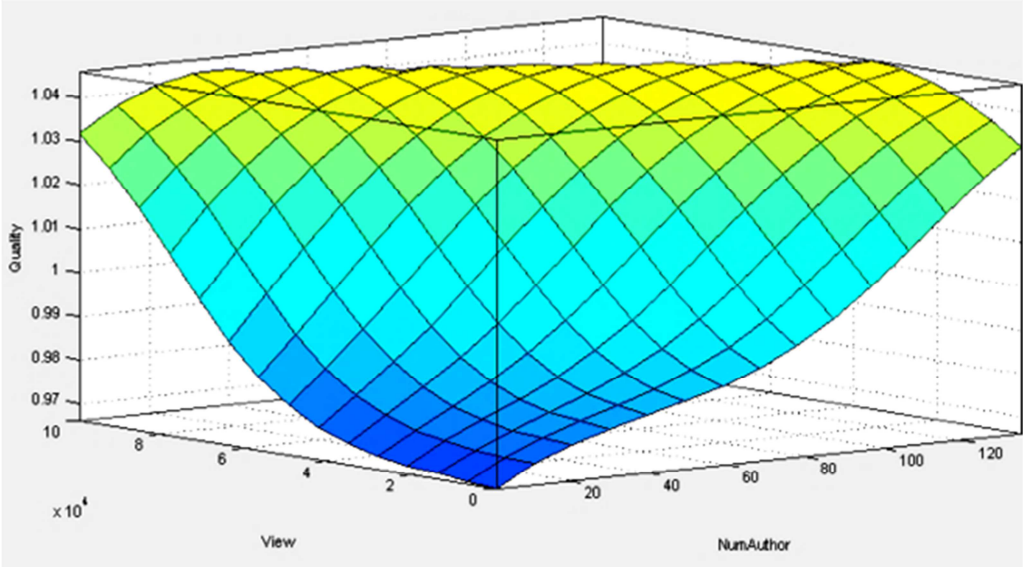


Figure 13B. *The increase in view and Num Authors variables cause an increase in articles quality.*

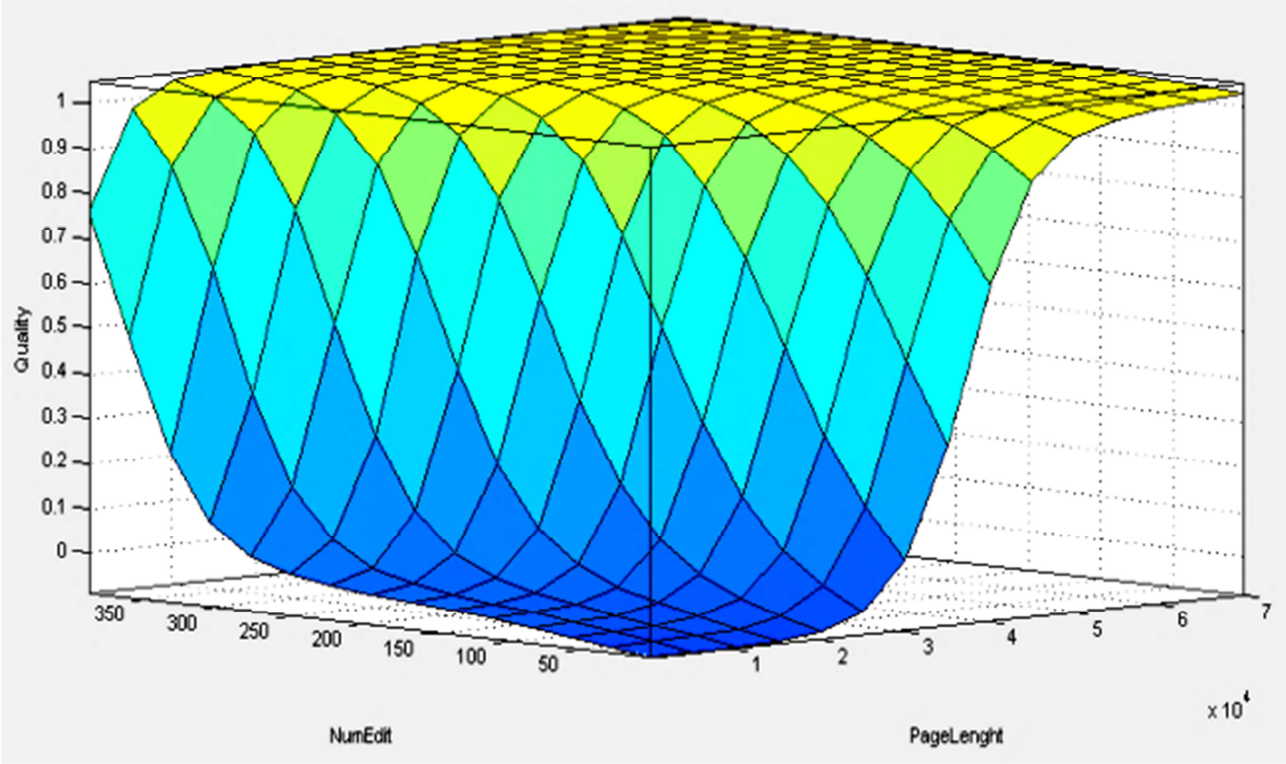


Figure 13C. *The increase in Num Edits and Page length variables cause an increase in articles quality.*

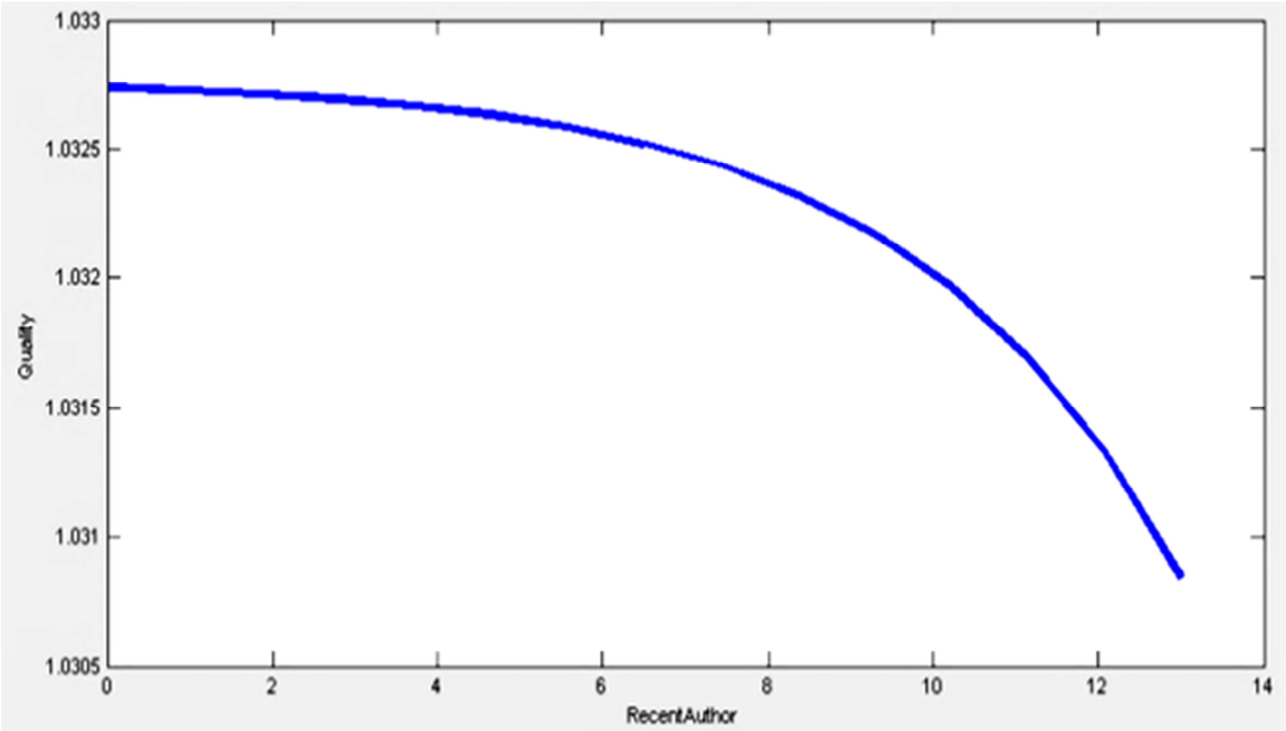


Figure 14. *Increasing the number of Recent Authors variable can lead to decrease the quality of articles.*

Table 1. *Compared two methods used in this research.*

Method	Rules	MFs		Type Of MF	Train Err		Chk Err	
		In	out		Epoch=0	Ep=600	Ep=0	Ep=600
(I)Decision tree & ANFIS	13	19	13	Trap	0.34	0.27	0.33	0.28
(II) Subtractive Clustering	5	35	5	Gaussian	0.31	0.27	0.32	0.29

5. Summery and Conclusion

In this study, the variables related to the articles listed in featured, good, cleanup and deletion Persian Wikipedia were collected. As it was noted in Section 4.1, due to the similarity of selected featured and good articles, these two variables have been studied and integrated under high quality articles. The collected data were entered into the Weka software, and then J48 decision tree has mined 13 rules of data relationships. At this part, the results had shown that 89.7% of the articles were classified correctly by decision tree. Confusion Matrix demonstrated that the decision tree separated high quality and deletion articles without error (shown in Fig. 4). Obtained rules to implement by ANFIS were entered into fuzzy toolbox of Matlab, linear relationship between inputs and outputs was established by ANFIS. In part 4.3, the variables for classification and implementation via Subtractive Clustering method were entered into Matlab software. 5 clusters were allocated to each variable, finally, articles were classified in three levels, including high quality, cleanup needed and deletion. As observed in the Figure 13, that the increased independent variables would cause increased articles quality. However, the figure 14, would indicate that increasing the number of recent authors can lead to decrease the quality of articles. Since when a paper was located as the group of cleanup or deletion articles, authors and editors try more to eliminate the article from those lists, Whereas in the case of articles selected as good and featured articles and their sublime position within encyclopedia recognized, subsequently, the authors of these articles have not been increased in the short term.

Table 1 has compared two methods used in this research, in Subtractive Clustering method, the number of rules reduced, while Membership functions and the error are increased in comparison to the first method. Consequently, the research query is answered and the objectives are fulfilled. Considering, the fact that classification of Wikipedia articles implemented user's opinion, it can be concluded that the system which is designed in separation or judgment articles is similar to users and site managers, style of thinking. By applying such a system, encyclopedia articles can be more quickly classified, and essays that might damage and discredit the can be promptly removed or revised validity Wikipedia and users trust quickly removed or revised. According to the results achieved by Stvilia et al. [51] in 2009 which showed that different communities have different models for measuring quality, the researchers of this paper have recommended such research to be done in other languages of Wikipedia encyclopaedia and its results utilized for raising the quality of encyclopaedia articles. The purpose of this study was to find the variables that would affect the quality of the Persian Wikipedia encyclopedia articles. It also aims at designing and implementation of a system to judge the Persian Wikipedia articles in three quality levels of high quality, cleanup needed, and deletion similar to users and administrators.

References

[1] Wikipedia. Persian Wikipedia, The Free Encyclopedia. Persian Wikipdia2013. https://en.wikipedia.org/wiki/Persian_Wikipedia (Persian Version).

[2] wikipedia. Statistics daily of persian wikipedia. Persian Wikipedia 2013. https://en.wikipedia.org/wiki/Wikipedia:Statistics (Persian Version).

[3] Liu J and Ram S. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. ACM Trans Manage Inf Syst. 2011; 2: 1-23.

[4] Kittur A and Kraut RE. Harnessing the wisdom of crowds in wikipedia: quality through coordination. Proceedings of the 2008 ACM conference on Computer supported cooperative work. San Diego, CA, USA: ACM, 2008, p. 37-46.

[5] Priedhorsky R, Chen J, Lam SK, Panciera K, Terveen L and Riedl J. Creating, destroying, and restoring value in wikipedia. Proceedings of the 2007 international ACM conference on Supporting group work. Sanibel Island, Florida, USA: ACM, 2007, p. 259-68.

[6] Wilkinson DM and Huberman BA. Cooperation and quality in wikipedia. Proceedings of the 2007 international symposium on Wikis. Montreal, Quebec, Canada: ACM, 2007, p. 157-64.

[7] Ullah N. ANFIS BASED MODELS FOR ACCESSING QUALITY OF WIKIPEDIA ARTICLES. Computer Engineering. Dalarna University, 2010.

[8] Wikipedia. Wikipedia: Protection policy. 2013. https://en.wikipedia.org/wiki/Wikipedia:Protection_policy (Persian Version).

[9] Wikipedia. Policies and guidelines. 2013. https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines (Persian Version).

[10] Wikipedia. Manual of Style. 2013. https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style (Persian Version).

[11] Wikipedia. Wikipedia: WikiProject Albums. 2013. http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Albums.

[12] Walraven A, Brand-Gruwel S and Boshuizen HPA. How students evaluate information and sources when searching the World Wide Web for information. Comput Educ. 2009; 52: 234-46.

[13] Lim S. How and why do college students use Wikipedia? J Am Soc Inf Sci Technol. 2009; 60: 2189-202.

[14] Lucassen T and Schraagen JM. Propensity to trust and the influence of source and medium cues in credibility evaluation. J Inf Sci. 2012; 38: 566-77.

[15] Stvilia B, Twidale MB, Smith LC and Gasser L. Information quality work organization in wikipedia. J Am Soc Inf Sci Technol. 2008; 59: 983-1001.

[16] Geiger RS and Halfaker A. When the levee breaks: without bots, what happens to Wikipedia's quality control processes? Proceedings of the 9th International Symposium on Open Collaboration. Hong Kong, China: ACM, 2013, p. 1-6.

[17] Rowley J and Johnson F. Understanding trust formation in digital information sources: The case of Wikipedia. J Inf Sci. 2013; 39: 494-508.

[18] Noč M and Zumer M. The completeness of articles and citation in the Slovene Wikipedia. Program. 2014; 48: 53-75.

[19] Lih A. Wikipedia as Participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource. Proceedings of the 5th International Symposium on Online Journalism. 2004, p. 16-7.

[20] Stein K and Hess C. Does it matter who contributes: a study on featured articles in the german wikipedia. Proceedings of the eighteenth conference on Hypertext and hypermedia. Manchester, UK: ACM, 2007, p. 171-4.

[21] Wöhner T and Peters R. Assessing the quality of Wikipedia articles with lifecycle based metrics. Proceedings of the 5th International Symposium on Wikis and Open Collaboration. Orlando, Florida: ACM, 2009, p. 1-10.

[22] Saengthongpattana K and Soonthornphisaj N. Thai Wikipedia Quality Measurement using Fuzzy Logic. The 26th Annual Conference of the Japanese Society for Artificial Intelligence. Japan2012, p. ROMBUNNO. 4M1-IOS-3C-1.

[23] Anderka M, Stein B and Busse M. On the evolution of quality flaws and the effectiveness of cleanup tags in the English Wikipedia. Wikipedia Academy. 2012; 2012.

[24] Xiao K, Li B, He P and Yang X-h. Detection of Article Qualities in the Chinese Wikipedia Based on C4.5 Decision Tree. In: Wang M, (ed.). Knowledge Science, Engineering and Management. Springer Berlin Heidelberg, 2013, p. 444-52.

[25] Chai K, Hayati P, Potdar V, Chen W and Talevski A. Assessing post usage for measuring the quality of forum posts. Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on. 2010, p. 233-8.

[26] Saravanan N, Cholairajan S and Ramachandran KI. Vibration-based fault diagnosis of spur bevel gear box using fuzzy technique. Expert Syst Appl. 2009; 36: 3119-35.

[27] Omid M. Design of an expert system for sorting pistachio nuts through decision tree and fuzzy logic classifier. Expert Syst Appl. 2011; 38: 4339-47.

[28] Jalali A and Mahmoudi A. Pistachio nut varieties sorting by data mining and fuzzy logic classifier. International Journal of Agriculture and Crop Sciences (IJACS). 2013; 5: 101-8.

[29] Chiu SL. Fuzzy Model Identification Based on Cluster Estimation. Journal of Intelligent and Fuzzy Systems. 1994; 2: 267-78.

[30] Yager RR and Filev DP. Approximate clustering via the mountain method. Systems, Man and Cybernetics, IEEE Transactions on. 1994; 24: 1279-84.

[31] Mathworks. Fuzzy Logic Toolbox: User's Guide (R2014a). 2014, http://www.mathworks.com/help/pdf_doc/fuzzy, pp. 2_109, 2_150, 2_156-158, 2_160-161.

[32] Gaur V, Soni A, Bedi P and Muttoo SK. Comparative Analysis Of ANFIS And ANN For Evaluating Inter-Agent Dependency Requirements. International Journal of Computer Information Systems and Industrial Management Applications. 2014; 6: 23-34.

[33] Yuan X, Khoshgoftaar TM, Allen EB and Ganesan K. An application of fuzzy clustering to software quality prediction. Application-Specific Systems and Software Engineering Technology, 2000 Proceedings 3rd IEEE Symposium on. 2000, p. 85-90.

[34] Wei L-Y, Chen T-L and Ho T-H. A hybrid model based on adaptive-network-based fuzzy inference system to forecast Taiwan stock market. Expert Systems with Applications. 2011; 38: 13625-31.

[35] Malhotra R and Sharma A. A neuro-fuzzy classifier for website quality prediction. Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on. 2013, p. 1274-9.

[36] Afshoon R, Harounabadi A and Mir Abedini J. Assessment and Validating the Quality of Educational Web Sites using Subtractive Clustering. International Journal of Computer Applications. 2014; 98: 42-7.

[37] Wikipedia. Cleanup. 2013. https://en.wikipedia.org/wiki/Wikipedia:Cleanup (Persian Version).

[38] Saunders MN, Saunders M, Lewis P and Thornhill A. Research methods for business students, 5/e. Pearson Education India, 2011, pp. 237-240.

[39] Wikipedia. Deletion policy. 2013. https://en.wikipedia.org/wiki/Wikipedia:Deletion_policy (Persian Version).

[40] Wikipedia. Good articles. 2013. https://en.wikipedia.org/wiki/Wikipedia:Good_articles (Persian Version).

[41] Wikipedia. Good article nominations. 2013. https://en.wikipedia.org/wiki/Wikipedia:Good_article_nominations (Persian Version).

[42] Wikipedia. Wikipedia: Featured articles. Persian Wikipedia2013. https://en.wikipedia.org/wiki/Wikipedia:Featured_articles (Persian Version).

[43] Wikipedia. Wikipedia: Featured article criteria. Persian Wikipedia2013. https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria (Persian Version).

[44] Wikipedia. Wikipedia article traffic statistics. 2013. http://stats.grok.se.

[45] Wikipedia. Wiki ViewStats. 2013. http://tools.wmflabs.org/wikiviewstats2

[46] Wikipedia. Glossary. 2013. http://en.wikipedia.org/wiki/Wikipedia:Glossary.

[47] Jang J-SR and Sun C-T. Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice-Hall, Inc., 1997, pp. 73-74.

[48] Sivanandam SN, Sumathi S and Deepa SN. Introduction to Fuzzy Logic using MATLAB. Springer-Verlag New York, Inc., 2006, pp. 123-124.

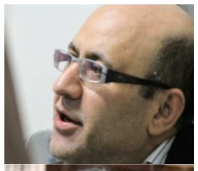
[49] Witten, Frank and Hall. Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition. 2011, pp. 410.

[50] Bouckaert RR, Frank E, Hall M, et al. WEKA Manual for Version 3-6-2. Hamilton, New Zealand: University of Waikato, 2011, pp. 21-22.

[51] Stvilia B, Al-Faraj A and Yi YJ. Issues of cross-contextual information quality evaluation—The case of Arabic, English, and Korean Wikipedias. Library & Information Science Research. 2009; 31: 232-9.

Biography

SeyedTaha Seyedsadr was born in Trier, Germany. He achieved a BSc in Telecommunication Engineering from *Islamic Azad University (IAU)* in Tehran. Currently, he is a master student in Information Technology Management at *IAU*. Mr. Sadr has been productive as a Research Expert in the *Graduate School of the Environment and Energy* at *the Science and Research Branch, IAU in Tehran* since 2008. His research interests include Discrete Process Signals, Digital Image Processing, Expert System, Data Mining, Web Mining, Fuzzy Systems, Neural Networks and Wavelet.



Mohammadali Afsharkazemi is an Associate Professor in the Faculty of *Management* at the *Tehran Central Branch, Islamic Azad University*. He received a PhD in Industrial Management from *Science and Research Branch, IAU*. His main research activities are would involve OR Theory, Opinion Mining and Web Content Mining, Intelligent Operation Management and Industrial Strategy setting, Industrial Automation & control, Fuzzy systems, Neural Networks and Genetic Algorithms.

Hashem Nikoomaram is a Professor in the Faculty of *Management and Economics at the Sciences and Research Branch, Islamic Azad University,* Tehran. A PhD in Commercial Management, he has been appointed as the dean of the Management and Economics faculty since 2004. Prof. Nikoomaram serves as the Director-in-Charge in Management Accounting Journal and Editor-in-chief of Financial Knowledge of Securities Analysis Journal. He also collaborates as an active member in several editorial boards. His preferable areas of interest in research fields are considered to be Manufacturing Systems, Supply Chain Management, Management Information Systems, Strategic Management, International Marketing Strategy.