Identification of Authors' Profiles in Wiki Using Quiz Game Based on Clusters Analysis Techniques

https://doi.org/10.3991/ijet.v14i09.10139

Mouna Boulaajoul (✉), Noura Aknin
Abdelmalek Essaadi University, Tetouan, Morocco
Mouna.blj@gmail.com

Abstract—The risk of inaccurate information found in a wiki-based site such as Wikipedia is high. In fact, it is a public site, so it is editable by anyone who can enter inaccurate information through malice or ignorance without any kind of control.

When entering or editing an article, it is recommended to have a clear style without mistakes. So, it is essential to get rid of the mediocre style and respect the rules of writing.

Even so, it is very remarkable that there are abuses at the style due to the malformed language coming even from people belonging to a more cultured social rank. Therefore, the need to identify the wiki authors' profiles has become paramount.

The idea of this article is to offer a quiz game in the intention to classify the language level of wiki authors by using data mining techniques to make groups where each group gets a significant result that we should analyze.

Keywords—WEKA, Data mining, Quiz game, Clustering, Author profile, Unity 3D.

1 Introduction

It was found that despite the relevance of the Wikipedia system and the rigorous monitoring applied by its Wikipedian administrative community, it remains an unreliable system according to the latest surveys conducted and shows several limitations [1]. The human factor has an impact on the validity and reliability of the information in a Wiki system and has a remarkable responsibility on the credibility of the published content.

Our work will focus on this subject and we will propose a quiz game to deduce the credibility of wiki authors and their content.

A Quiz game is a form of entertainment in which the authors compete in answering questions. The research study aims to identify relevant data from a quiz game answered by authors and the appropriate data mining methods for deduction of authors' profiles, such as in [2].

To reach our goal, the realization of this work will be mainly based on data mining techniques. Among the chosen algorithms, we find clustering and K-means algorithm.

Regarding the software, we will use Weka workbench which is a collection of learning algorithms and data preprocessing tools. It includes all the algorithms that can be used for clustering.

The entries of the database that we should analyze are: the number of right responses, the number of the incorrect responses and the timing duration taken to answer questions in this quiz game.

The software used in developing this quiz game is Unity. It is a multiplatform 3D software used for creating games and 3D / 2D objects on mobile platforms, web and consoles. It is available as a free license for many of its features and is supported by a large community of users and developers.

2 Related Work

In [3], the need to have a system that allows to associate texts with a reputation index of its author (in terms of quality of content) has become vital. The Wikitrust system is characterized by certain entries that will be tinged with a color (more or less bright orange for example) to alert readers to the fact that the information has not been verified to ensure its reliability or credibility.

The color code is based on an automatic update algorithm, taking into account an author's reputation (based on his / her previous contributions, and the lifetime of an article (It is assumed that the longer it persists, the more reliable it is likely to be) [3].

The idea of this project is very important. However, this tool is used only in the websites installed in the MediaWiki platform.

A study by a group of researchers [4] tried to highlight the importance of distinguishing the role of different types of users by searching the determinants of the value of their contribution.

Interests and resources are considered micro inputs for the group. Interests regroup individual motivations that highlight why people contribute, while resources can be associated with knowledge or skills that individuals can use in their contributions into collective wikis (Oliver et al., 1985) [4].

There are two types of contributors depending on their editing types:

- The substantive changes where the user can add new information
- The non-substantive changes where the user corrects the spelling of already existing content

The results indicated that if the contributors with high depth of resources and interest focus on substantive changes, they would make a valuable content. On the other hand, if the contributors with high breadth of resources and interests focus on non-substantive changes, they would provide a valuable contributions [4].

3 Methodology

3.1 Proposal quiz game

The proposed quiz game analyzes performances and behaviors of the authors. Furthermore, this quiz let to see the level of understanding in grammar of all authors in a distractive and relaxed way (gaming) [5]. On the other hand, the major difference with other tests of knowledge is that the participant shouldn't develop words to answer. All of that offers us a possibility to analyze the data issued and treat it using the pre-mentioned data mining techniques.

This quiz may be in the form of a simple questionnaire: we will develop a quiz that consists of a set of questions in the area of grammar. This questionnaire aims to deduce the language grammar levels of the authors.

The figure 1 shows the schema of our developed application. The established quiz game has multiple questions and every question has four options where one answer must be selected by the contributor. After that, the quiz shows the next question randomly and so forth until the end of the quiz.

According to the number of good and bad answers and the time consumed to answer the questions in the quiz, a level of the contributor will be defined and deduced.

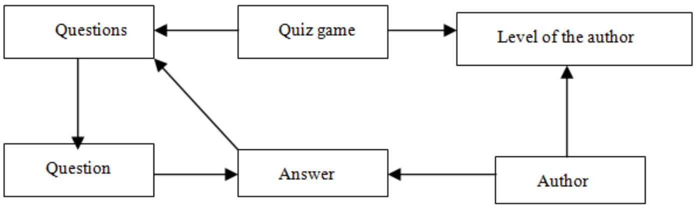


Fig. 1. Schema of our quiz game

3.2 Authors type

The authors in Wikipedia edit the articles by adding many contents or just making a little change and spend relatively more time rewriting and reordering pages.

In this study we will focus on the style and the rules of writing, because it is very remarkable that there are abuses at the style caused by malformed language and this can occur even with persons belonging to a high cultured social rank.

There are many types of authors according to their grammar level. Our effort will concentrate on extracting the authors groups from the results issued by them on the quiz game.

We will be able in the end to obtain a number of authors divided in six groups by using the data mining clustering:

- Advanced language level of authors, with a short duration to answer
- Advanced language level of authors, but need some guidance

- Intermediate language level of authors, but need some guidance
- Intermediate language level of authors, and choose responses quickly
- Beginner language level of authors, and have chosen responses randomly
- Beginner language level of authors, and make effort to understand

3.3 Database architecture

The quiz game will be equipped by a database, see the class diagram in Figure 2. The information like number of good choices, bad choices and duration describe all the behaviors of the author. In addition, personal information such as pseudo, name, last name and email will be saved in this database. The table "QUIZZ" contains all the questions, and the four choices shown in the game including the response. All of this data will help to analyze author's performances by feeding the k-means algorithm in order to group the authors according to their language performances.

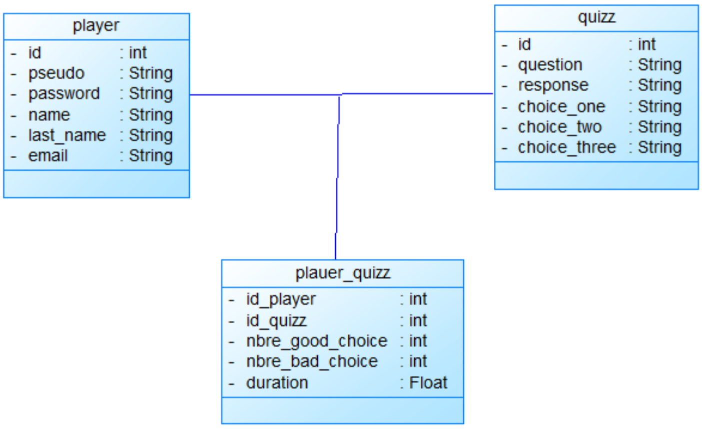


Fig. 2. Quiz game class diagram

3.4 Quiz game engine

Game engine is a tool available for game developers to code and make out a game quickly and easily, as stated in [6].

The engine can be used to create both three-dimensional and two-dimensional games. The software Unity has the particularity of using a mono compatible script editor (C #), as well the Unity engine is a multi-platform gaming engine (smartphone, Mac, PC, video game consoles and web)[7], All information about author and game should be stored in a Database server (MYSQL), see Figure 3.

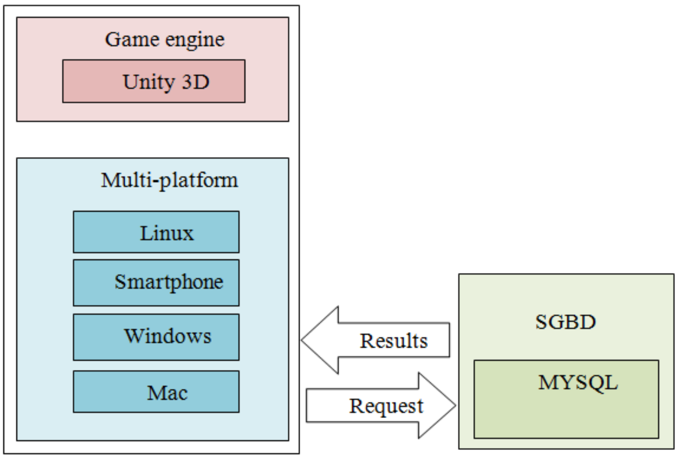


Fig. 3. Quiz game architecture

Unity is one of the most widespread in the video game industry. The software has declined in four formulas, classified in order of price:

- The free license "Personal"
- The paid license "Plus"
- The paid license "Pro"
- The paid license "Enterprise"

In our case, we will use the "Personal" version.

3.5 Development of the quiz game

When an author wants to edit content in the wiki or to add a new one, a quiz game will be shown in front of the authors. It is crucial to respond to all questions before having access to editing the wiki application.

The main objective of the proposed quiz game is to extract the language level of each author as stated in [8]. This author will try to select the right response.

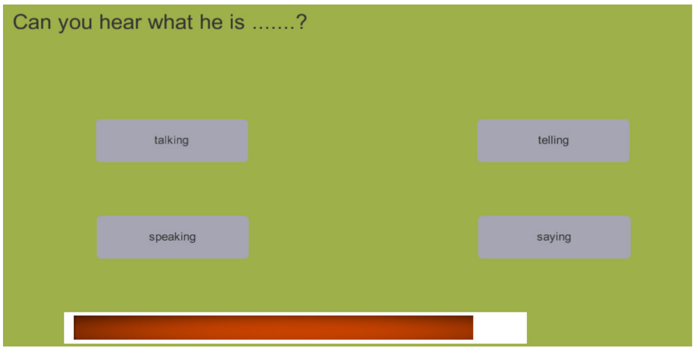
In our study, we have prepared ten questions, that may let us to know the exact grammar level of the author in multiple topics in the grammar area:

- Question 1: Can you hear what he is ...? (Talking, telling, speaking).
- Question 2: I can't come to school … I have an important appointment. (So, because, unless).
- Question 3: That smells good! What… (Are you cooking?, do you cooking?, do you cook?, are you cook?).
- Question 4: It was the first time he … anything so spicy (had eaten, has eaten, had).
- Question 5: She's wearing a … dress (long black beautiful, black long beautiful, beautiful long black, long beautiful black).

- Question 6: How long have they… there? (been waiting, waited, waiting, been waited).
- Question 7: If I had more time, I … do more exercise (want to, 'm going to, would, will).
- Question 8: Winters there … be really cold sometimes (can, might, could, may).
- Question 9: Take a sandwich with you … you get hungry later (if, in case, so as not to, when).
- Question 10: That wasn't a good idea - you … thought about it more carefully (must have, have to, should have, ought to have).
- Question 11 : The film … by Quentin Tarantino (was direct, directed, did directed, was directed).

It is not necessary to have just ten questions: It can be more if we want to know more information and data about the profile of the authors.

The application is developed in Unity 3D and the language C Sharp. The Figure 4 shows a question interface that will be presented to the author. The quiz was taken by twenty three users who tried to answer the questions correctly by selecting the right answer according to their knowledge.



Can you hear what he is ....?

Fig. 4. Game question interface

3.6 WEKA data mining software

The software WEKA (Waikato Environment for knowledge Analysis) is a free open source data mining framework. It is available under the GNU General Public License. The WEKA workspace is a collection of machine learning algorithms for data mining tasks, and contains algorithms for data analysis and a collection of visualization tools, as stated in [9].

This software is written in the Java language and contains a GUI (Graphical User Interface) to interact with data files and produce visual results. We will load data by the User Interface Chooser WEKA, such as in [10]. The Figure 5 shows this interface.

Waikato Environment for Knowledge Analysis
Version 3.6.13
(c) 1999 - 2015
The University of Waikato
Hamilton, New Zealand

Fig. 5. WEKA startup screen

Like in [11], the use of clustering techniques in this quiz game is vital to classify authors into groups based on the duration and the number of the good and the bad choices made by them. If authors make a good choice, the reward will be the gain of one point. However, in the opposite case, the number of bad choices will be incremented taking in consideration the duration taken to choose the answer.

The clustering results have revealed six forms of authors' participation, who interact with the game according to their experiences.

The proposed quiz game has been developed on Unity 3D. Thus, it needs only a web browser or an android to be run.

3.7 Clustering algorithm in WEKA

Clustering is considered the most important unsupervised learning algorithm; so every problem of this kind deals with finding a structure in a collection of data. A loose definition of clustering is the process of organizing objects into groups whose members are similar in some way. Then, a cluster is a collection of objects which are similar between them and dissimilar to the objects belonging to other clusters.

The algorithm used to partition data is: k-means algorithm. This algorithm is used to classify the given data objects into different clusters through k centroids for each cluster as stated in [12].

The WEKA K-Means algorithm uses Euclidean distance to compute distances between clusters.

The Euclidean distance for multi-dimensional points (1) measures the distance from an instance "x" to the average point centroid "y", where "n" is the number of attributes [13].

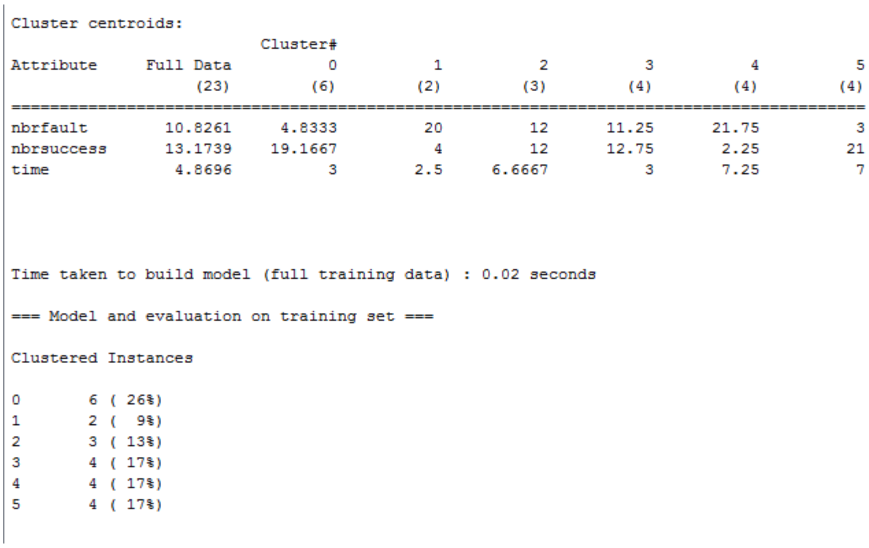
$$\boldsymbol{d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}} \tag{1}$$

We use clustering techniques to classify wiki authors' profiles into predefined groups based on the number of wrong and right responses and the duration of game issued by the authors.

In the next part, we have the results obtained in our study. The prepared data was then put through the data mining process. The K-means algorithm was used in this step. The number of six clusters was determined.

4 Results

We have found six types of groups in our study. We extracted from WEKA framework the number of instances as results that are, then, classified according to the kind of group as shown in Figure 6.



```
Cluster centroids:
                         Cluster#
Attribute     Full Data         0          1          2          3          4          5
                   (23)       (6)        (2)        (3)        (4)        (4)        (4)
========================================================================================
nbrfault        10.8261    4.8333         20         12      11.25      21.75          3
nbrsuccess      13.1739   19.1667          4         12      12.75       2.25         21
time             4.8696         3        2.5     6.6667          3       7.25          7




Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 26%)
1       2 (  9%)
2       3 ( 13%)
3       4 ( 17%)
4       4 ( 17%)
5       4 ( 17%)
```

Fig. 6. Results of clustering of the data with 6 clusters in WEKA

The clusters are classified according to the type of group. For example:

- The Cluster 0 has the value of the "nbrfault" less than the value of the "nbrsuccess" and the value of the time is small.
- The Cluster 1 has the value of the "nbrfault" more than the value of the "nbrsuccess" and the value of the time is small.
- The Cluster 2 has the value of the "nbrfault" equals to the value of the "nbrsuccess" and the value of the time is great.
- The Cluster 3 has the value of the "nbrfault" approximately equals to the value of the "nbrsuccess" and the value of the time is small.
- The Cluster 4 has the value of the "nbrfault" more than the value of the "nbrsuccess" and the value of the time is great.
- The Cluster 5 has the value of the "nbrfault" less than the value of the "nbrsuccess" and the value of the time is great.

The use of the data mining technique in the quiz game provides us with significant findings and may lead to conclude the profile of each author according of his/her answers and the duration to answer all the questions in the Quiz game. From the obtained profile, we can deduce the language level of this author and allocate them in a pre-defined group.

All of the extracted results are detailed in Table 1.

Table 1. kind of group of each cluster number

Cluster	nbrfault	nbrsuccess	time	Cluster's group
Cluster 0	4.83333	19.1667	3	Advanced language level of authors, with a short duration.
Cluster 1	20	4	2.5	Beginner language level of authors, and have chosen responses randomly.
Cluster 2	12	12	6.6667	Intermediate language level of authors, but need some guidance.
Cluster 3	11.25	12.75	3	Intermediate language level of authors, and choose responses quickly
Cluster 4	21.75	2.25	7.25	Beginner language level of authors, and make effort to understand
Cluster 5	3	21	7	Advanced language level of authors, but need some guidance.

These obtained data show the profiles of authors and allow us to classify all of them in pre-defined groups. Consequently, once an author edits or modifies a content in wiki, the reader of this content will be able to know the profile of this author. In this way, the reader will be aware of the quality of the information in every article he/she tries to read.

5 Conclusion

Knowing the users' profiles in any application will be highly crucial. In this context, the proposed quiz game could be used in many areas to extract the profiles of the authors, especially, in the E-Learning domain which is of high importance and a real recourse in the learning system nowadays.

This situation has made the check and verification of the quality of the data published online, especially in a wiki system, very difficult since no real means to control the credibility of the authors and their sources have been set out. By having an idea about the author's reputation and the level of their profile, the reader can avoid a lot of incorrect and misleading content on the web.

This study uses data mining in the field of the Wiki. The analysis of this situation is different from the usual data mining studies. Cluster analysis and K-means analysis were used as data mining techniques. The steps of the data mining process were carried out and explained in detail.

The clustering revealed six groups of authors' profile:

- Advanced language level of authors, with a short time to answer.
- Advanced language level of authors, but need some guidance.
- Intermediate language level of authors, but need some guidance.
- Intermediate language level of authors, and choose responses quickly.
- Beginner language level of authors, and have chosen responses randomly.
- Beginner language level of authors, and make effort to understand.

6 References

[1] S. Javanmardi, Y. Ganjisaffar, C. Lopes and P. Baldi, "User contribution and trust in Wikipedia," 2009 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing, Washington, DC, 2009, pp. 1-6. https://doi.org/10.4108/ICST.COLLABORATECOM2009.8376

[2] B. Sbihi, K. E. El Kadiri,N. Aknin, "Towards an implementation of the concepts of Elearning 2.5 through one group of ten Master's learners : Case of the UML course," *International Journal of Emerging Technologies in Learning*, vol. 8, Issue 4, Aug. ,pp. 68-73, 2013. https://doi.org/10.3991/ijet.v8i4.2920

[3] B. T. Adler, L. De Alfaro, and I. Pye, "Detecting wikipedia vandalism using WikiTrust: Lab report for PAN at CLEF 2010," CLEF 2010 LABs and Workshops, Padua, Italy, 22-23 September 2010.

[4] S. J. Zhao, K. Z. K. Zhang, C. Wagner, and H. Chen, "Investigating the determinants of contribution value in Wikipedia," *International Journal of Information Management*, vol. 33, no. 1, pp. 83–92 , February 2013. https://doi.org/10.1016/j.ijinfomgt.2012.07.006

[5] H. Bicen, S. Kocakoyun, "Perceptions of Students for Gamification Approach : Kahoot as a Case Study," *International Journal of Emerging Technologies in Learning*, vol. 13, no. 2, pp. 72-93, 2018. https://doi.org/10.3991/ijet.v13i02.7467

[6] Y. Daineko, M. Ipalakova, D. Tsoy, Y. Yelgondy, A. Shaipiten, "Using new technologies to support physics course for secondary schools," EDULEARN18 Proceedings, 2018, pp. 3559-3562. https://doi.org/10.21125/edulearn.2018.0918

[7] G. T. A. Kusuma, I. M. A. Wirawan, I. K. R. Arthana, "Virtual Reality for Learning Fish Types in Kindergarten," *International Journal of Interactive Mobile Technologies*, vol. 12, no. 8, pp. 41-50, 2018. https://doi.org/10.3991/ijim.v12i8.9246

[8] H. A. AbuAlsaad, R. AlTaie, "Using Big Data Technology for Prediction of Quiz Difficulty Level in E-learning Systems," Iraqi Journal of Information Technology, 2018, V .8, N. 4.

[9] N. Sharma, A. Bajpai, R. Litoriya, "Comparison the various clustering algorithms of weka tools," International Journal of Emerging Technology and Advanced Engineering, vol. 2, ISSN 2250-2459, 2012.

[10] S. Z. Erdogan, M. Timor,"A data mining application in a student database," *Journal of aeronautics and space technologies*, vol. 2, pp. 53-57, 2005.

[11] K. Anagnostou, M. Maragoudakis, "Data Mining for Player Modeling in Videogames," 13th Panhellenic Conference on Informatics, 2009, pp. 30 – 34. https://doi.org/10.1109/PCI.2009.28

[12] Y. Zhang, J. Lin and Hui Zhang, " A Hierarchical Teaching Mode of College Computer Basic Application Course Based on K-means and Improved PSO Algorithm," *International Journal of Emerging Technologies in Learning*, vol. 11, Issue 10, pp. 53-58, 2016. https://doi.org/10.3991/ijet.v11i10.5909

[13] N. K. Visalakshi , K. Thangavel, "Impact of Normalization in Distributed K-Means Clustering," *International Journal of Soft Computing*, vol. 4, pp. 168-172, 2009.

7 Authors

Mouna Boulaajoul: Got the Master's degree in Computer Systems and Network Engineering in 2011 from Abdelmalek Essaadi University in Tangier, Morocco.

She is a PhD Student in Computer Science in Laboratory of Computer Science, Operational Research and Applied Statistics in Abdelmalek Essaadi University in Tetuan, Morocco. She is the responsible of IT Department in the court of appeal of Tangier. Her current research is focusing on Web 2.0, Data minning, author's credibility in wiki-based system and the quality of the content wiki.

Noura Aknin: Professor of Electrical & Computer Engineering at Abdelmalek Essaadi University since 2000. She received PhD degree in Electrical Engineering in 1998. She is the Head of Research Unit Information Technology and Modeling Systems. She is the Co-founder of the IEEE Morocco Section since November 2004 and a member of several IEEE societies. She is R&D project manager/member related to new technologies and their applications. She was a chair of several conferences and has been involved in the organizing and the Scientific Committees of several international conferences held worldwide dealing with e-learning, Mobile Networks, Social Web and information technologies. Her research interests focus mainly on mobile and wireless networks, Social web and e_learning. She is and author of several papers on e-learning, mobile and wireless communications, Web 2 applications.

Moreover, she has supervised several Ph D and Masters Theses. aknin@ieee.org

Article submitted 2019-01-10. Resubmitted 2019-02-28. Final acceptance 2019-03-05. Final version published as submitted by the authors.