The Role of the Clusters Analysis Techniques to Determine the Quality of the Content Wiki

https://doi.org/10.3991/ijet.v14i01.9074

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

Abstract—The online sources of the web, since years, are an extraordinarily important base of information and knowledge. Indeed, the web is one of the best access point to any type of information. For the users who want to share their knowledge, the wiki system is a powerful tool.

Nevertheless, any system has its limits. The investigation on the contributions performance of individual contributors is yet unexplored because it is partly related to the design of wikis, which is considered for collaborative work. Consequently, this has made the assessment and evaluation of individual contributions a hard task.

In this research, we will attempt to emphasize the significance of distinguishing the relevant articles based on the opinions of contributors and their contributions. In this way, we will focus on the utilization of data mining using clusters analysis and k-means algorithm techniques.

Keywords—web 2.0, Wiki, data mining, cluster Analysis, k-Means algorithm, discretization, WEKA.

1 Introduction

The Research on the credibility of online contributions is still unexplored due to the complexity of the architecture of wikis, which is designed for shared work.

The web 2.0 is described by its adaptability and interactivity [2], particularly at the interfaces that enable users to cooperate and express their opinions [9] on the wiki such as rating, comments and social bookmarks.

This wiki interface can be considered as part of a web 2.0 approach [10], if it uses in a privileged way the following modules:

Bookmarking wiki module: Users save links to web pages they find interesting.

Rating Wiki module: This is another practical method used to obtain relevant information according to the rating average obtained for each article.

Comment Wiki Module: Comments give users the opportunity to start a debate in case of disagreement.

In this context, we will try to extract the information from the contributors' interactions to evaluate the pertinence of the articles distributed online. To reach this purpose, we will use the techniques of clustering to form groups of articles in an automatic way.

2 Cluster analysis

Clustering allows to make groups of data. We can create a specific number of groups depending on our needs [4]. One defining benefit of clustering is that every attribute in the data set will be used to analyze the data [3]. However, a major disadvantage of using clustering is that we are required to know ahead of time how many groups we want to create. It might take several steps of trial and error to determine the ideal number of groups to create. In our study we need eight groups of articles, it is classified in this way, see Table 1:

Table 1. Kind of group

Kind of group	Group number
The articles are more relevant The subjects are more important A great debate	Group 0
The articles are less relevant The subjects are more important A great debate	Group 1
The articles are more relevant The subjects are less important A great debate	Group 2
The articles are more relevant The subjects are more important A small debate	Group 3
The articles are less relevant The subjects are less important A great debate	Group 4
The articles are less relevant The subjects are more important A small debate	Group 5
The articles are more relevant The subjects are less important A small debate	Group 6
The articles are less relevant The subjects are less important A small debate	Group 7

Basically, clustering can be the most useful data mining method we can use. It can take our entire set of data and turns it into groups, from which we can make some conclusions. The math behind the method is complex and this is the reason behind the use of WEKA software.

3 Data Mining Framework

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.

This software is written in the Java language and contains a GUI (Graphical User Interface) to interact with data files and produce visual results. It has a general API in such a way that we can set up WEKA, like any other library, in own applications as automated server-side data-mining tasks. There are two means to load data. In our work, we will focus on loading data by the User Interface Chooser WEKA [7].

4 Methodology

In this study, we will focus on the application of data mining to get relevant articles based on the opinions of users and their contributions, using cluster analysis and discretization to get relevant results with an error equal to zero.

4.1 Clustering Data Mining Method

A cluster is a collection of objects, which are similar to each other, and different from the objects belonging to other clusters [6]. A major inconvenient of using clustering is the need to know ahead of time how many groups we want to create.

In our study, we need eight groups of articles. Clustering can quickly take the entire set of data and turn it into groups, from which we can easily make some conclusions. The math behind the method is complex that is why we have opted for the use of WEKA.

We create a specific number of groups based on our needs. The selection of attributes has a very important role to play in improving the results of clustering, see Figure 1.

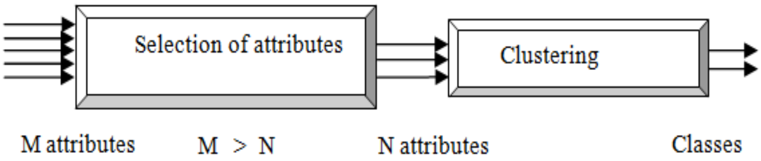


Fig. 1. Selection of attributes to improve the classification of a cluster

4.2 Discretization in weka

We can turn a numerical attribute into a categorical one by some sort of discretization. So, the process of discretization is an essential task of the data preprocessing, because the data is transformed in a set of intervals [6]. This involves dividing the range of values into subranges called bins that are more relevant to make the computation process goes faster.

We have two methods to make this, based on our need [6].

The Equal-width binning algorithm (1) divides the range of values into K bins of the same size. The best way of determining K bins is by looking at the type of groups we want to have. The width of bin is:

$$width\ of\ bin = \frac{(\max - \min)\ value}{k} \tag{1}$$

The Equal-frequency algorithm divides the range of values into N groups where each group contains approximately the same number of values.

We load the data of thirty rows, that we need to study, into WEKA by executing an SQL query. In our case, we select three attributes: “avg_rating”, “sum_bookmarking” and “sum_comment”.

For example, we extract twelve values of the attribute “avg_rating”:[5 , 2 ,3, 3.2 , 1 , 1.5 , 2 , 4 , 3.75 , 3.125 , 2 , 0]

These values are splitted into two intervals in the code, without using any methods of the discretization, because the min value is always fixed on the number zero and the max value is fixed on the number five. We create two intervals as below:

width = (5 – 0)/2 = 2.5
bins: [- , 2.5], (2.5, +] .
bin 1 : 2 , 1 , 1.5 , 2 , 2 , 0
bin 2 : 5 , 3.2 , 4 ,3.75 , 3.125 ,3

We use the Equal-frequency algorithm on the attributes: “sum_bookmarking” and “sum_ comment”.

We extract as an example twelve values of the attribute “sum_ comment”:

Sum_comment: [0, 2, 7, 10, 14, 15, 17, 20, 22, 35, 67, 54]

We calculate a value halfway between the values on either side of the boundary, for example :(17 + 15)/2 = 16

bins: [-,16], (16, +]
bin 1 : 0, 2, 7, 10, 14, 15
bin 2 : 17, 20, 22, 35, 54, 67

These results present the more and less commented articles.

We obtain these results of dividing our data set of thirty rows into scales:

[Avg_rating_interval1] = A1 = [- , 2.5]
[Avg_rating_interval2] = A2 = (2.5, +]
[Sum_bookmarking_interval1] = B1 = [- , 55]
[Sum_bookmarking_interval2] = B2 = (55, +]
[Sum_comment_interval1] = C1 = [- , 53]
[Sum_comment_interval2] = C2 = (53, +]

4.3 Simple k-means algorithm in WEKA

The k-means algorithm is used specifically in unsupervised learning to partition instances into k groups, often called clusters. k must be defined previously as stated in [4].

The function used to measure the distance between instances is called Euclidean distance.

The Euclidean distance (2) measure the distance from an instance X to the average point centroid Y, where n is the number of attributes [5]. The Euclidean distance for multi-dimensional points is:

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

If there are nominal attributes the distance (3) is:

$$d(x_i, y_i) = \begin{cases} 0 \; if \; x_i = y_i \\ 1 \; if \; x_i \neq y_i \end{cases} \quad (3)$$

The algorithm starts with defining the initial cluster. Thereafter, the distance is calculated from each instance to cluster centroid by the Euclidean distance. For example, we have an instance A, if this instance is nearer to the cluster X than other clusters, then, we rank the instance A in cluster X in step 1.

We continue our treatment with other instances but we cannot be sure that each instance has been assigned to the appropriate cluster. So, we compare each instance's distance to its own cluster centroid and to that of the opposite cluster. If the instance's distance is nearer to the centroid of the opposite cluster, it will be relocated to it.

The iterative relocation would now continue from this new partition until no more relocation occurs. The iteration stops by choosing the latest partitioning as the final cluster solution.

The attribute data of an instance is scaled to fit into a specific range. So, the Euclidean distance is not dominated by any attribute as stated in [5].

K-Means algorithm handles a numerical and categorical attributes and normalizes numerical attributes when computing distance as stated in [1].

The Min-Max Normalization (4) transforms a value A to B, which fits in the range [C, D]. It is given by the formula below:

$$B = \left(\frac{A - \min value\; of\; A}{\max value\; of\; A - \min value\; of\; A}\right) \times (D - C) + C \quad (4)$$

In the range (0, 1) we get this formula (5):

$$B = \left(\frac{A - \min value\; of\; A}{\max value\; of\; A - \min value\; of\; A}\right) \quad (5)$$

We click on the Cluster tab to create clusters. We click Choose and select Simple K-Means algorithm from the choices that appear. We are interested in the Num Clusters field, which tells how many clusters we want to create. Let us change the default value to number eight.

At this point, we are ready to run the clustering algorithm and WEKA calculates thirty rows of data with eight data clusters [8].

5 Results

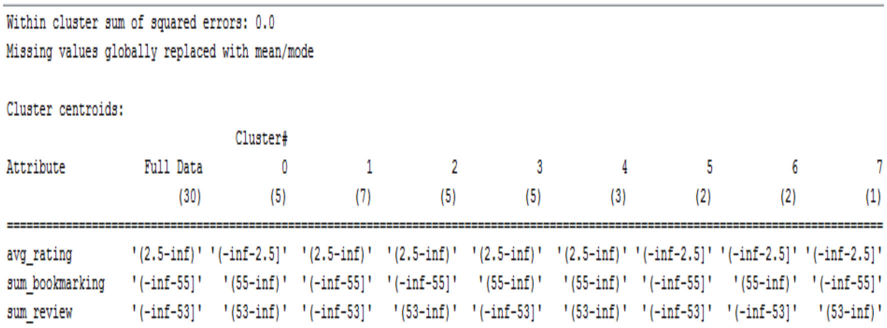
In the next part we have the results obtained in our study; the first result is about to deduce the quality of the content wiki.

5.1 Wiki reliability

The data is typically broken up into the testing data that contains thirty instances.

In the end, we get these results with eight cluster; the output tells us how each instance comes together in a cluster. We have eight clusters, in each cluster we have a number of instances, see Figure 2.

The error is the distance for each point to the nearest cluster, to get SSE we square the distance between each member of the cluster and its centroid and sum them. We notice that the sum of squared errors is 0, because we transform continuous to nominal value. So, for nominal attributes, the distance is 0 if both instances in question have the same value and 1 if they are different.



```
Within cluster sum of squared errors: 0.0
Missing values globally replaced with mean/mode

Cluster centroids:
                           Cluster#
Attribute        Full Data        0            1            2            3            4            5            6            7
                      (30)       (5)          (7)          (5)          (5)          (3)          (2)          (2)          (1)
==============================================================================================================================
avg_rating     '(2.5-inf)' '(-inf-2.5]'  '(2.5-inf)'  '(2.5-inf)'  '(2.5-inf)'  '(2.5-inf)' '(-inf-2.5]' '(-inf-2.5]' '(-inf-2.5]'
sum_bookmarking '(-inf-55]'  '(55-inf)'  '(-inf-55]'  '(-inf-55]'   '(55-inf)'   '(55-inf)'  '(-inf-55]'   '(55-inf)'  '(-inf-55]'
sum_review     '(-inf-53]'  '(53-inf)'  '(-inf-53]'   '(53-inf)'  '(-inf-53]'   '(53-inf)'  '(-inf-53]'  '(-inf-53]'   '(53-inf)'
```

Fig. 2. Results of clustering of the data with 8 clusters in WEKA

In our study, we extract from the results the values of instances that are classified according to the kind of group. we have eight groups of articles, it is classified in this way, see Table 2.

Table 2. Kind of group of each cluster number

Kind of groups	Instance	Number of Instances	Interval	Cluster number
The articles are more relevant The subjects are more important A great debate	7,13, 20, 23,25	5	'(2.5-inf)' '(55-inf)' '(53-inf)'	3
The articles are less relevant The subjects are more important A great debate	1,8,9,15, 16 , 18 , 19	7	'(-inf-2.5]' '(55-inf)' '(53-inf)'	1
The articles are more relevant The subjects are less important A great debate	3,11,17,26, 28	5	'(2.5-inf)' '(-inf-55]' '(53-inf)'	2
The articles are more relevant The subjects are more important A small debate	4,5,6, 10, 29	5	'(2.5-inf)' '(55-inf)' '(-inf-53]'	0
The articles are less relevant The subjects are less important A great debate	24	1	'(-inf-2.5]' '(-inf-55]' '(53-inf)'	7
The articles are less relevant The subjects are more important A small debate	14 , 22 , 27	3	'(-inf-2.5]' '(55-inf)' '(-inf-53]'	4
The articles are more relevant The subjects are less important A small debate	2,30	2	'(2.5-inf)' '(-inf-55]' '(-inf-53]'	5
The articles are less relevant The subjects are less important A small debate	12 , 21	2	'(-inf-2.5]' '(-inf-55]' '(-inf-53]'	6

6 Conclusion

The idea of this study is to use the data provided by the collaborators to create an intelligent wiki platform. This wiki application is considered as belonging to web 2.0 approach, because it is based on the collaborative sociological aspect between collaborators.

We developed three modules in our application:

- **Social bookmark:** It shows how a site is perceived and help users to store their favorite links. Namely, in a wiki site, users save wiki articles they find useful and important.
- **Rating:** It is an explicit way of getting feedback on how the user like this article, his advantage is the information provided is quantifiable, and can be exploited.
- **Comment:** The users express their opinions through comments left after reading a wiki article, especially, in case of disagreements.

The study utilizes data mining to discover kind of wiki articles, moreover this technique demonstrates more varied and significant findings, and may lead to show the quality of the content wiki.

Cluster analysis is used as data mining technique. Indeed, the selection of attributes has a very important role to improve the result of clustering.

K-means analysis process is carried out and explained in detail and consists of creating a specific number of groups, depending on our needs. K-Means algorithm handles a numerical and categorical attributes and normalizes numerical attributes when computing distance. The function used to measure the distance between instances, is the Euclidean distance measure.

The experimentation shows the advantages of using discretization techniques to get results with the value of the error is zero, because it turn a numerical attribute into a nominal/categorical one, by using some sort of discretization.

7 References

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

[2] B. Sbihi, K. El Kadiri and N. Aknin, "Towards a participatory E-learning 2.0 based on the use of Vwiki tool," International Journal of Innovation and Applied Studies, vol. 1, pp. 178-185, 2012.

[3] 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.

[4] 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.

[5] N. Kaur, K. Kumar, "Normalization Based K-means Data Analysis Algorithm," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, ISSN 2277 128X, 2016.

[6] I. H. Witten, E. Frank, Data mining, United States of America: Elsevier, 2005.

[7] R. Bouckaert, E. Frank,M. Hall, R. Kirkby, P. Reutemann, A. Seewald, D. Scuse, WEKA Manual for Version 3-6-13 , Hamilton New Zealand : University of Waikato, 2015.

[8] Dr. Sudhir B. Jagtap, Census Data Mining and Data Analysis using WEKA, (ICETSTM – 2013) International Conference in "Emerging Trends in Science, Technology and Management-2013", Singapore, 2013, Retrieved from https://arxiv.org/abs/1310.4647

[9] Z. Itahriouan, N. AKNIN, A. Abtoy and K. E. El Kadiri." Building a Web-based IDE from Web 2.0 perspective". International Journal of Computer Applications 96(22):46-50, June 2014.

[10] Boubker Sbihi, Kamal Eddine El-Kadiri, and Noura AKNIN, Towards a mixed methodology based on the groups of Vblog and Vwiki to the collaborative E-learning, ARPN Journal of Systems and Software, VOL. 3, NO. 3 March-April 2013 ISSN 2222-9833.

8 Authors

Mouna Boulaajoul: received the Master 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. Here current research interest is Web 2.0, Data Minning, author's credibility in wiki 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 eLearning. 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.

Article submitted 20 June 2018. Resubmitted 24 July 2018. Final acceptance 06 August 2018. Final version published as submitted by the authors.