A clustering approach to infer Wikipedia contributors' profile

Shubham Krishna
IIT(ISM)
Dhanbad, India
shubhamkrishna@am.ism.ac.in

Romain Billot
IMT Atlantique & Lab-STICC
Brest, France
Romain.Billot@imt-atlantique.fr

Nicolas Jullien
IMT Atlantique & M@rsouin-LEGO
Brest, France
Nicolas.Jullien@imt-atlantique.fr

ABSTRACT

Recent studies have improved our knowledge about the different types or profiles of online contributors, from casual to very involved ones, through focused people. But they use very complex methodologies, making their replication by the practitioners limited. We show on both Romanian and Danish wikis that using only the edit and their distribution over time to feed clustering techniques, allows to build these profiles with good accuracy and stability. This suggests that light monitoring of newcomers may be sufficient to adapt the interaction with them and to increase the retention rate.

CCS CONCEPTS

•Human-centered computing → Empirical studies in collaborative and social computing;

KEYWORDS

Online communities; clustering; user profile; Wikipedia

ACM Reference format:
Shubham Krishna, Romain Billot, and Nicolas Jullien. 2018. A clustering approach to infer Wikipedia contributors' profile. In *Proceedings of The 14th International Symposium on Open Collaboration, Paris, France, August 22–24, 2018 (OpenSym '18),* 5 pages.
DOI: http://10.1145/3233391.3233968

1 INTRODUCTION & PAPER'S GOAL

In open, online communities different profiles of contributors exist, regarding their level of involvement and focus [1, 9, 12, 13]. Better detecting these profiles is important for such project managers, to better adapt their response to newcomers contributions, and to improve the retention rate [7]. Recent studies [2, 15, 16] have strongly improved our knowledge of those different profiles, from casual to very involved ones, through focused people. But the are mostly on English projects (English Wikipedia for instance). They also use very complex methodologies (qualitative-quantitative mix, with a high workload to manually codify/characterize the contributions).

If these studies could be extended beyond the English speaking projects, it is not sure that 1) they could go farther in terms of precision in the description of the different profiles, 2) the contributors would invest their time to manually create the dataset of coded contributions these methods require[1]. We wonder here whether observing only the number of edits over time makes it possible to elicit the same profiles as with those more complex methods.

OpenSym '18, Paris, France
 978-1-4503-5936-8/18/08…$15.00
DOI: http://10.1145/3233391.3233968

2 RESEARCH METHODOLOGY

We studied the Danish and Romanian Wikipedia, two mid-size projects. We used WikiDAT[2] for extraction of data from the dumps. The study was limited to those contributors who had contributed more than 100 edits (irrespective of whether the edits made were minor or major)[3]. We removed those contributors who were either robots, contributed only in a single month or contributed anonymously. There were such 171 contributors in the Romanian Wikipedia and 274 contributors in the Danish Wikipedia.

Construction of the variables

Our goal was to use simple activity measures based only on the edits and their distribution over time. Contributors are likely to be grouped in terms of volume, intensity (focus) or duration of the activity. Starting with 12 initial features, we obtained, after studying the correlation matrix, a short list of 6 features, described in Table 1.

Ratio measures how massively the contributors contributed during their entire period of contribution and incorporates

[1] Wikipedia is a good example of this problem. The development of artificial intelligent tools is very active for the 'big' Wikipedias, but slower for the smaller ones. Their tuning requires human contribution, not always easy to find. See the ORES project, on detecting the quality of the edits for an example of this difficulty https://www.mediawiki.org/wiki/ORES.
[2] http://glimmerphoenix.github.io/WikiDAT/
[3] We discuss this number in the Conclusive Discussion Section.

Table 1: Description of the variables

Variable	Description
Ratio	Ratio between the number of edits & the number of days a contributor has been on Wikipedia from the very first edit
Mean_gap	The average gap between 2 consecutive posts measured in months
Max_gap	The maximum gap between any two consecutive posts measured in months.
Num_cons	The number of pairs of consecutive months with contributions
Mean_Month	Average of the monthly number of edits
SD	Standard Deviation among the month average edits value

the relationship between the number of edits and the number of days. SD provides information about the variations in the contributions made during these months. The features Ratio, Mean_Month & SD taken together evaluate the quantity and deviation of the contributions made by the contributors. Both Mean_Gap & Max_Gap inform about how often the contributors get active and how long they can quit the community before coming back. Num_cons tells about how many times the contributors have contributed successively for two consecutive months. For example, if a contributor made edits in January 2011 and February 2011, the count is increased by 1. It is a measure of the regularity of contributors over time.

Statistical methods

We used the Romanian Wikipedia to calibrate the methods and come up with a first groups interpretation, and the Danish Wikipedia to check the group correspondence across different datasets. A contribution of this article is to provide this double checking in terms of cluster validation. A two-stage cluster analysis was performed:

(1) We did a hierarchical clustering based on the features described in the Table 1 with the *hclust* function of the R platform. We used was the Ward distance, adapted to quantitative features [5].
(2) We used partitioning algorithms as alternative clustering methods in order to select the final typology. The contributors were clustered using a k-medoids clustering algorithm called PAM (Partitioning Around Medoids), from the R package *cluster*. PAM is based on the search for k representative objects or medoids among the observations of the data set. It is more robust than the k-means algorithm, especially for the initialization [10].

Different typologies were formed for k ranging in the interval selected in step 1. Results were consistent with those obtained from step 1. The optimal number of cluster was selected with validation technique such as the silhouette index [8] (intra vs. inter cluster inertia).

To assist the interpretation of the resulting clusters, we carried out Principal Component Analysis (PCA) in order to project the data onto a small number of dimensions (a combination of the initial variables) [14]. Three dimensions were enough to explain almost 90% of the data variability. In addition to PCA, ANOVA analysis and Tukey statistical tests helped to determine the significant variables within each cluster. This ensures a full and robust interpretation of clustering results.

3 RESULTS

The dendrogram of Hierarchical clustering suggests an interval between 2 and 10 clusters for the optimal number of clusters k (Figure 1). Figure 2 depicts the evaluation results

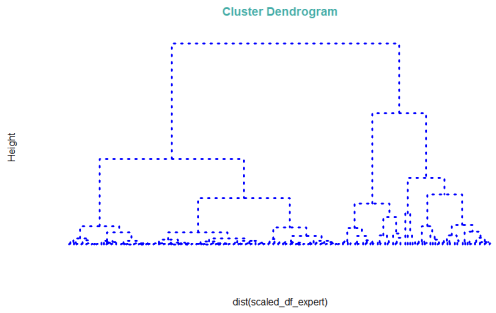

Figure 1: Cluster Dendrogram - Romanian Wikipedia

in four clusters of contributor's contribution behavior in Romanian Wikipedia. This number was validated afterward with the Danish Wikipedia. The distribution of the contrib-

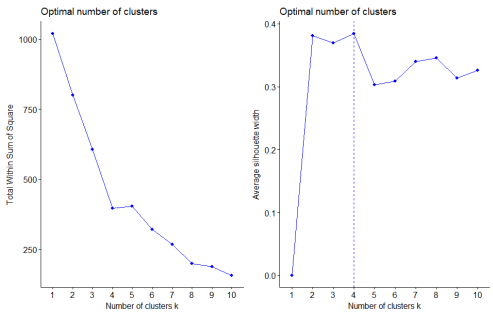

Figure 2: Cluster Validation Plots - Romanian Wikipedia

utors in the clusters for both the Wikipedias are in Table 2. Regarding cluster interpretation, a PCA with three principal components explains almost 90% of the total variance. Figure 3 projects the labeled contributors onto these three dimensions. For both wikis, the first dimension (PC1) is correlated with the volume of the activity (ratio, mean number of edits) with a relative intra-cluster variability. Dimension 2 (PC2) relates to the periods of inactivity (the gaps - the correlation is negative). Dimension 3 (PC3) refers to the variable Num_cons, mainly, so to the notion of regularity. The number of articles involved has been added as an illustrative variable, to better link our findings to the descriptions realized by [3, 4].

Table 2: Size of Clusters

Wikipedia	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Romanian	25	92	48	6
Danish	45	144	61	24

This leads to the following contributors profiles:

- Cluster 1: contributors 'on a mission'. the typical user of this cluster has some periods of activity but not that intense, separated with periods of zero-activity (a characteristic of this cluster). The illustrative variable 'number of article they worked on' shows a focus on a small number of articles, sometimes just one in a period of activity.
- Cluster 2: basic, or 'casual' contributors. Their activity is never particularly intense. Their involvement is less regular, in terms of level of contribution per month, than Cluster 3 contributors. They are not particularly focused on a subset of articles.
- Cluster 3: regular contributors. The activity is above the average (even if not that much) and the most regular among all groups, and regular (number of consecutive months of presence)
- Cluster 4: top contributors. They have huge activity ratios. Nevertheless, this cluster contains higher variability than others.

These interpretations are confirmed by unidimensional boxplots distributions (Figure 4 and Figure 5, still in Appendix).

4 CONCLUSIVE DISCUSSION

Are simple measures of contributing activities over time with data reduction techniques enough to detect the different profiles of contributors? At least on Wikipedia, we have detected the focused workers (Cluster 1), the casual workers (Cluster 2), and the regular workers (Clusters 3 & 4, and, even discriminated between those, the very involved (Cluster 4). Simple data reduction techniques such as clustering and PCA have provided a comparable level of information as more refined approaches, such as Non-parametric Hidden Markov Clustering models of profiles [15]. Although the method looks simple, the combination of hierarchical clustering with PAM (which is, however less used, much more robust than k-means), plus a PCA for the interpretation, is quite original. Moreover, the work was deepened with a refined clustering validation process (a benchmark of 13 cluster quality indexes was made with the R package ClValid). The methodology makes our findings less sensitive to other datasets than most of the existing studies can be.

There is a strong and very applicable result from this work. The highlighted profiles can be identifiable early in the history of involvement, suggesting that light monitoring of newcomers may be sufficient to adapt the interaction with them and increase the retention rate. For instance, while for Cluster 2 to 4 members, the Wikipedia teahouse initiative may be efficient[4], for Cluster 1 participants, the interaction has to be done at article level, as proposed by [6], as those participants are very specific in their interest. And Cluster 1 profiles seem quite easy to detect early in their life as Wikipedian (as soon as the first period of contribution, focused on one or two articles).

However, we made a strong hypothesis by focusing on the contributors with more than 100 edits. If a potential application is to increase the users retention rate, it would be relevant to pay a special attention to the small contributors with less than 100 edits, and design retention strategies for them. We did the same analyses on contributors with more than 50 edits, as it is the definition of an editor (or 'Wikipedian'), a least in the French project[5]. Our results were the same, suggesting that profiles can be determined with very few edits. The exact limitation is to be explored in the future, as a trade-off between earlier profile detection and increasing noise (quality issues and uncertainty).

It would be interesting for a managerial point of view to run the same analyses on Arabic, or Thai, or Hindi Wikipedias, in a word any non-occidental Wikipedias, to see if the same profiles exist. The simplicity of the method in term of data collection, and its language-specificity insensitivity make this transfer much easier than for other methods.

There is room for improvement of the light monitoring of the contributors' behavior and their profiling, too. Research would gain to extend this off-line clustering towards dynamic techniques (dynamical adaptation of the clusters as new contributors join the community) to develop a dynamic

[4]The Wikipedia Teahouse, is a virtual 'safe place' where newcomers are welcomed and coached by old-timers. This program seems to increase the retention rate, see [11]

[5]Only People having more than 50 non-anonymous edits are allowed to vote for the administrators, for instance.

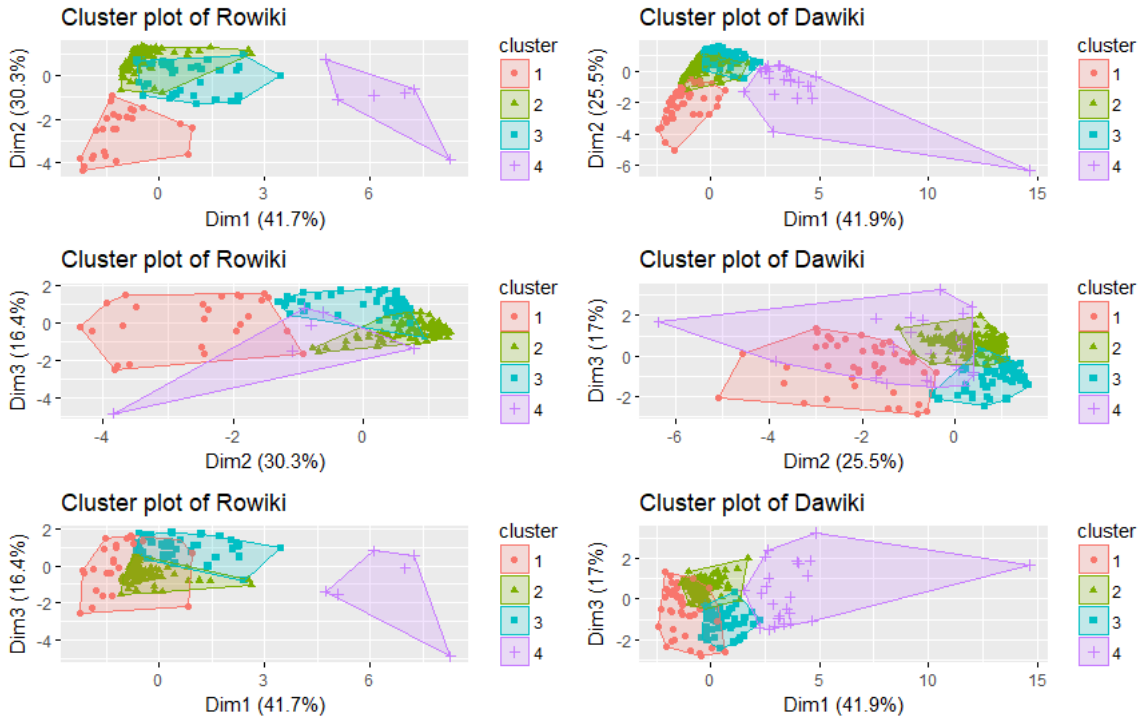

Figure 3: PCA Analysis with projection of the four clusters

decision support tool for nesting and assistance. Methods such as Growing Neural Gas could be used for that goal.

REFERENCES

[1] Judd Antin, Coye Cheshire, and Oded Nov. 2012. Technology-mediated contributions: Editing behaviors among new wikipedians. In *Proceedings (CSCW '12)*. ACM, 373–382.

[2] Ofer Arazy, Johannes Daxenberger, Hila Lifshitz-Assaf, Oded Nov, and Iryna Gurevych. 2016. Turbulent stability of emergent roles: The dualistic nature of self-organizing knowledge coproduction. *Information Systems Research* 27, 4 (2016), 792–812.

[3] Ofer Arazy, Hila Lifshitz-Assaf, Oded Nov, Johannes Daxenberger, Martina Balestra, and Coye Cheshire. 2017. On the" How" and" Why" of Emergent Role Behaviors in Wikipedia. In *Proceedings (CSCW'17)*. 2039–2051.

[4] Martina Balestra, Ofer Arazy, Coye Cheshire, and Oded Nov. 2016. Motivational Determinants of Participation Trajectories in Wikipedia.. In *ICWSM*. 535–538.

[5] Richard O Duda, Peter E Hart, and David G Stork. 2012. *Pattern classification*. John Wiley & Sons.

[6] Andrea Forte, Niki Kittur, Vanessa Larco, Haiyi Zhu, Amy Bruckman, and Robert E. Kraut. 2012. Coordination and beyond: social functions of groups in open content production. In *Proceedings (CSCW '12)*. ACM, 417–426.

[7] Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. 2013. Making Peripheral Participation Legitimate: Reader Engagement Experiments in Wikipedia. In *Proceedings (CSCW '13)*. ACM, 849–860.

[8] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. 2002. Cluster validity methods: part I. *Sigmod Record* 31, 2 (2002), 40–45.

[9] Corey Jackson, Carsten Østerlund, Veronica Maidel, Kevin Crowston, and Gabriel Mugar. 2016. Which Way Did They Go?: Newcomer Movement Through the Zooniverse. In *Proceedings (CSCW '16)*. ACM, 624–635.

[10] Leonard Kaufman and Peter J Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.

[11] Jonathan T Morgan and Aaron Halfaker. 2018. Evaluating the Impact of the Wikipedia Teahouse on Newcomer Retention. In *Proceedings (OpenSym'18)*. ACM.

[12] Sneha Narayan, Jake Orlowitz, Jonathan Morgan, Benjamin Mako Hill, and Aaron Shaw. 2017. The Wikipedia Adventure: Field Evaluation of an Interactive Tutorial for New Users. In *Proceedings (CSCW'17)*. ACM.

[13] Chitu Okoli and Wonseok Oh. 2007. Investigating recognition-based performance in an open content community: A social capital perspective. *Information & Management* (2007).

[14] Gilbert Saporta. 2006. *Probabilités, analyse des données et statistique*. Editions Technip.

[15] W. Wei, C. Liu, M. Y. Zhu, and S. A. Matei. 2015. A Non-parametric Hidden Markov Clustering Model with Applications to Time Varying User Activity Analysis. In *Proceedings (ICMLA'15)*. IEEE, 549–554.

[16] Diyi Yang, Aaron Halfaker, Robert E Kraut, and Eduard H Hovy. 2016. Who Did What: Editor Role Identification in Wikipedia.. In *ICWSM*. 446–455.

APPENDIX

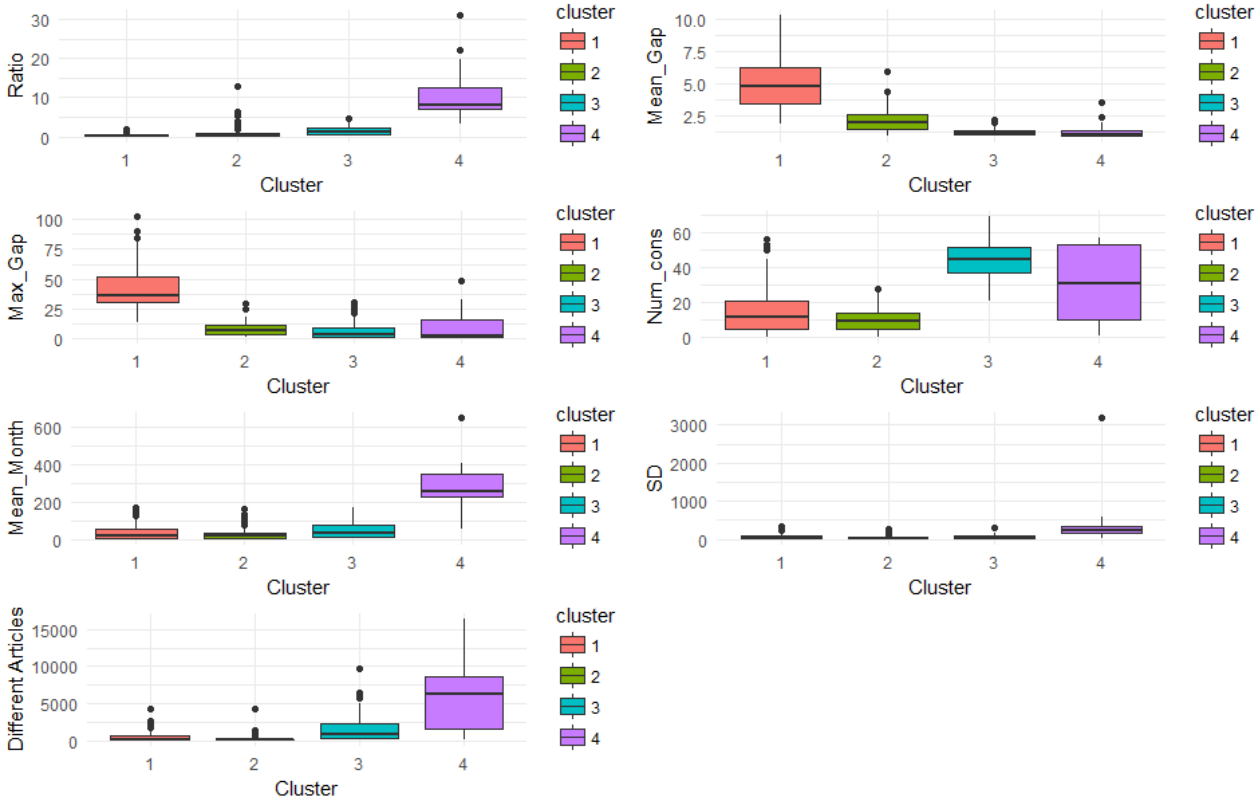

Figure 4: Boxplot of features distribution within each cluster for the Danish Wikipedia

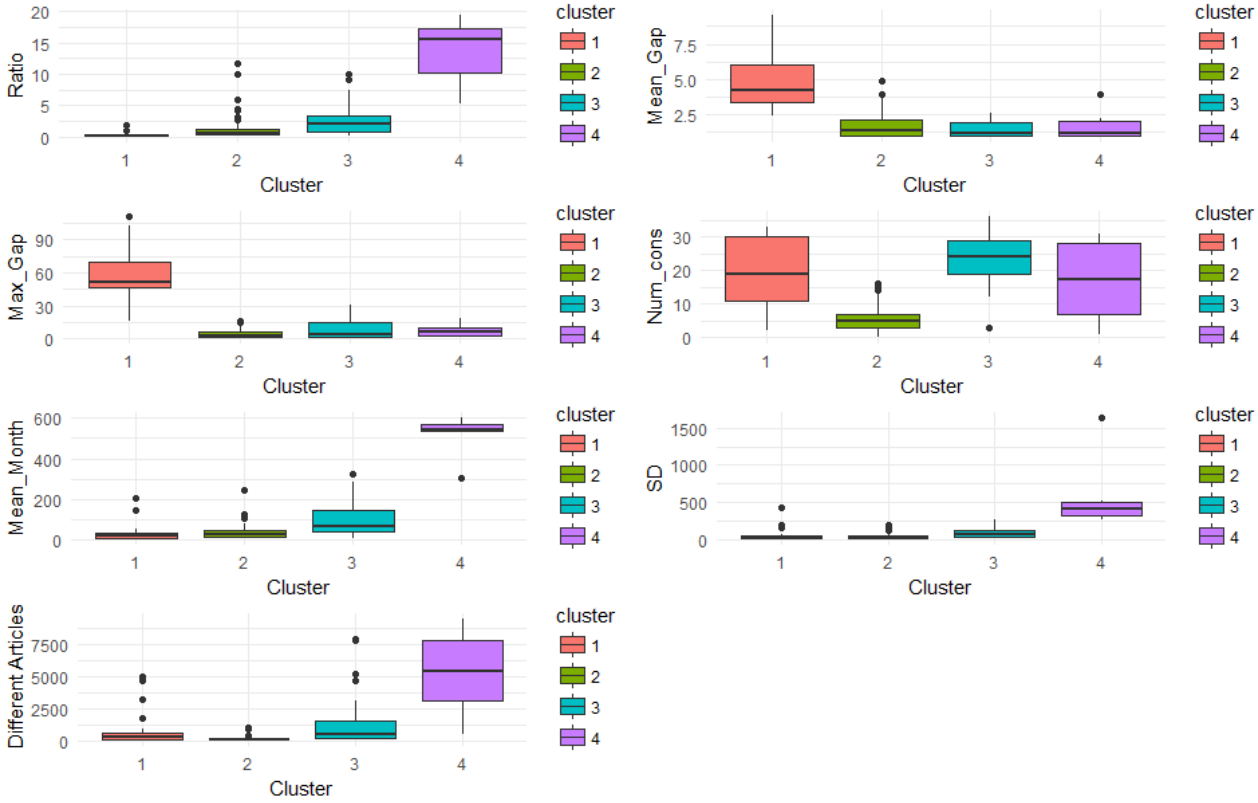

Figure 5: Boxplot of features distribution within each cluster for the Romanian Wikipedia