Mining team characteristics to predict Wikipedia article quality

Grace Gimon Betancourt, Armando Segnini, Carlos Trabuco, Amira Rezgui, Nicolas Jullien

Mining team characteristics to predict Wikipedia article quality

Grace Gimon Betancourt, Armando Segnini, Carlos Trabuco
Télécom Bretagne
FirstName.Name@telecom-bretagne.eu

Amira Rezgui, Nicolas Jullien
Télécom Bretagne-UBL, M@rsouin-LEGO
{Amira.Rezgui, Nicolas.Jullien}@telecom-bretagne.eu

ABSTRACT

In this study, we were interested in studying which characteristics of virtual teams are good predictors for the quality of their production. The experiment involved obtaining the Spanish Wikipedia database dump and applying different data mining techniques suitable for large data sets to label the whole set of articles according to their quality (comparing them with the Featured/Good Articles, or FA/GA). Then we created the attributes that describe the characteristics of the team who produced the articles and using decision tree methods, we obtained the most relevant characteristics of the teams that produced FA/GA. The team's maximum efficiency and the total length of contribution are the most important predictors. This article contributes to the literature on virtual team organization.

Keywords

Wikipedia, Epistemic community, Article Quality, Teaming

1. INTRODUCTION

Mobilizing hundreds or thousands of contributors, such as Linux and Wikipedia, volunteer online open projects aimed at creating new knowledge are viewed as the main source of generation of further new, innovative knowledge by and for firms. They are (virtual) epistemic communities, or task-oriented groups, bringing experts together around a common goal [27]. Online epistemic communities are specific in the sense that people do not necessarily know each other, they interact virtually, mainly via the Internet and data management tools such as Wiki systems (MediaWiki) or software version systems (Git). But the questions they raise are, in most of the cases, the same as for other epistemic communities : how do people organize themselves to produce knowledge ? What is the quality of the production ? What are the characteristics of the 'good' teams, those who produce 'good' knowledge, or knowledge of quality ? How can we define 'quality' ?

These questions are not new to the scientific literature. The critical mass theory of the construction of collective action [23], the theory analyzing the construction of the (knowledge) commons [14], and even closer to our question, studies of groups' creativity and efficiency [36], stress that these projects are made possible by the aggregation of various motivations and levels of involvement, but also various levels of competence and of intimacy among the members. However, the right balance between number and expertise, the size of the supposed critical mass is still a matter of debate [1].

In this study, we check whether it is possible to predict the quality of the article by looking at the same type of team characteristics in virtual organizations.

This assessment was done using Wikipedia as a case study for the characteristics of the contributors and of the articles, but also because of data availability. Wikipedia's contributors come from different academic or professional backgrounds, making it very hard to assess quality by the authority of the writers. Wikipedia has developed its own internal quality rating procedure, which relies on discussions and consensus building among reviewers and nominators who evaluate the candidate articles on their prose, lead, comprehensiveness, structure and style [2]. The highest ranked articles in terms of quality are named the 'Featured Articles' (FA), followed by the 'Good Articles' (GA). In this study we used Spanish Wikipedia dumps [3] of November 11, 2015. This prediction was achieved by using predictive data mining techniques [10], and more specifically those dedicated to large data sets since we used a dump of around 24 GB.

The rest of this paper is organized as follows : We discuss the question of measuring the quality of the article and the link to the team producing it in section 2. In section 3 we present our approach, which addresses the problem of unlabeled articles and the identification of the most important team characteristics to detect qualified articles (FA/GA). We present our main results in section 4 and discuss their implication for both theory and practice in section 5. Section 6 concludes with a discussion of the limitations and the possible extensions to this work.

2. BACKGROUND

2.1 What do measuring and predicting article quality mean ?

Nowadays, the quality of a common knowledge, especially of a Wikipedia article, is socially constructed and constantly evolving [34]. Wikipedia article quality may be evaluated using external measures such as Katz's criteria : purpose, authority, scope [17]. In

1. On this, and about the functioning of Wikipedia sub-part, or WikiProjects, see [29].
2. Featured Article Candidates : https://en.wikipedia.org/wiki/Wikipedia:Featured_article_candidates, Featured List Criteria : https://en.wikipedia.org/wiki/Wikipedia:Featured_list_criteria
3. http://Dumps.wikimedia.your.org

OpenSym'16 August 17-19, 2016 | Berlin, Germany

ACM ISBN
DOI : .
DOI: 10.1145/1235

this context, [18] discussed the difficulty of applying such criteria to Wikipedia and gave a proposal for quality evaluation criteria for such encyclopedia. To evaluate an article, we can use also internal measures, such as Wikipedia's own quality grading, which give categorization schema assigning articles to a set of 7 distinct categories : from 'Stub' (poor) to 'Featured' (very good) Articles. Wikipedia's article evaluation can focus also on the improvement path of the articles, in terms of time for bug correction, coping with damage or vandalism, etc. [6, 37], or on the process of production and IQ assurance patterns [34], or on revision efficiency [4]. Readers' experience may also be considered, as it may be the goal of an encyclopedia to provide the information needed to its users, with, again, various measures : user's perceived quality [12], with all the bias, it entails or external accessibility measures [22].

Finally, as stressed by [9], who proposed a list of the measures of an open online project success (open source software). There is a large variety and a more problematic non-convergence of the possible measures. To solve this problem, a point of view has to be chosen, which restrains, but also defines the scope of the analysis. Here, since we are dealing with the question of identifying the characteristics of a good team to produce good knowledge, we are taking the project management viewpoint, as exposed by [16] : these projects are (virtual) epistemic communities, or task-oriented groups, bringing experts together around a common goal [27], here the building of (new) knowledge available to other people (explicit, published online knowledge, being programs or encyclopedic articles). In that sense, we rely on two subsets of the literature about the detection of qualified articles : the one based on the composition of the articles and the other based on the characteristics of the team.

2.2 Good and Featured articles

Several metrics based on the article itself have been used to assess the quality of Wikipedia's articles. For instance, [33] explored the assessment of information quality of a Wikipedia Article measured by their defined metrics such as Completeness which depends on the number of internal broken links, the number of internal links, and first and foremost, the length of the article. Additionally, as stressed by [5], beyond length the other metrics are computing intensive, in terms of both information retrieval (article's history of revision) and data analysis.

[2, 3] proposed a methodology and the first results on the impact of team composition on article quality : as the subset of FA articles is small and as the quality of such articles is varied [20]. They rely on the external evaluation of the articles to find that having a small and a very committed team with strong local inequality improves the coordination and thus indirectly the quality, and having strong global inequality (people very invested in Wikipedia and peripheral contributors) improves the quality of the articles. As they stressed, this work may be extended to a bigger set of articles to be confirmed.

Currently, since this work, the rating of Wikipedia articles has improved and the automatic techniques to assess the quality of the non-labeled articles as well. One of the problems for labeling the articles in Wikipedia is that it is not only a question of assessing quality but also of having somebody who monitors this labeling. [19] showed that machine learning can be used to assess the quality of an article, by comparing its characteristics with those of a Featured Article. However, they mix information about the team (the editors) and of the article (Featured Article or not) on the same network, leading them to conclude that articles of quality are those edited by editors who have also edited Featured Articles, as done by [39].

As a conclusion, [18] proposed a list of 13 criteria to evaluate the quality of an article by looking at its content. The main criteria to assess this quality is the length of the article, even if the style of the article matters too [21]. We will follow their lead and focus on this simple measure to classify the articles.

2.3 Team characteristics as factors for a high-quality article

A set of team characteristics has been linked to performance in producing knowledge outputs of quality, beyond the literature. It is out of the scope of this article to check it extensively [5], and we will present here the main variables we used in our modeling. Most of them are taken from [38], for comparability reasons.

2.3.1 Efficiency, or reputation and experience

The high quality of an article is often related to many of contributions/contributors in Wikipedia. Some investigations cited by [30] explored German Wikipedia and showed that featured articles are not necessarily written by a huge number of people, but the most important is to be written by some contributors feeling personally responsible for the article, and thus involving themselves in their writing.

In the same context, [30] showed that the number of contributors' past contributions is an interesting measure of their efficiency and reputation, which positively impacts the quality of the present production (the article they are contributing to).

However, from a resource-based view, members who are involved in multiple projects may have less time for each project : the amount of time spent on one activity cannot be spent on another. Having contributors participating in too many concurrent projects may decrease the likelihood of obtaining high-quality contributions, and a high-quality article [38].

2.3.2 Tenure

In online communities, it is often the case that members who have been active for a long time tend to be more experienced than newcomers. These active members play a fundamental role in the community in terms of broadcasting knowledge, information and experience among the whole community. In this context, tenure, which is measured as the time that the individuals have spent in the community from the date that they made their first contribution, has been shown as a good predictor of performance. Existing literature posits a curvilinear relationship between tenure and performance [31]. When a newcomer joins a group, performance increases over time, as this participant acquires experience, accumulates new knowledge, develops skills, and becomes familiar with the new area and the rules of the organization. A similar effect of tenure is expected in the context of Wikipedia [38]. So, longer tenure allows for more experience, which helps increase productivity.

But after some years of effort, participants are more prone to a lassitude effect, which may lead to a decrease in productivity and performance [4].

All these works helped us refine our analysis and our goals. In this article, we propose 1) to classify the articles in terms of quality measured by their structure and proximity with Featured or Good

4. Such as the "revision score", developed by the Wikimedia Foundation, which measures the quality of a revision, https://meta.wikimedia.org/wiki/Research:Revision_scoring_as_a_service, which is inspired by, among others, [1]

5. The reader interested by these questions, may consult, in addition to [9], already cited for online communities, [35], for FLOSS, and [3], for Wikipedia, the review of the literature made by [24] on team effectiveness.

articles, via machine learning techniques, 2) to assess the characteristics of the teams which have produced such high-quality articles.

3. PROPOSED APPROACH

3.1 Presentation of our workflow

Our workflow consisted in obtaining the dump of the Spanish Wikipedia [6], and extracting the meta data about each article (regarding its structure, size, etc. and its contributors), with the WikiDAT tool [7]. As a starting point for this project we used only the page, people, and revision tables generated by WikiDAT, which correspond respectively to the data from the pages in Wikipedia (such as articles, subcategories, etc...), the data from the person who contributed to this page and the intermediary revision table to link these two. This table contained 2,782,466 articles. Spanish Wikipedia currently has over 1,200,000 articles related to the topics ; the rest of the articles contained in this table are redirection articles and disambiguation articles [8]. However, this tool does not give the labeling of the articles (Featured, Good, not labeled article). To get this label, we relied on the Wiki API developed by the Wikimedia Foundation [9].

Those labeled articles correspond to about 1% of the data, not because the other articles are not on the same level, but because, for most of them, the internal labeling process has not evaluated them. [39] discussed this point and proposed several techniques to classify unclassified articles. These techniques allowed us to predict the class of the already classified articles, but this did not change the problem of learning from the attribute what a good or a featured article is. We took from them to develop a simple approach based on proximity classifying the articles. As they pointed out the difficulty in differentiating the Featured Articles (FA) from the Good Articles (GA), we regrouped them into one single category (GA/FA). We applied a semi-supervised learning task from only positive and unlabeled data in order to label 99% of the data, with an equally good literature accuracy (see below).

Knowing the label of the articles we analyzed the characteristics of its contributors via decision tree techniques. The "Contributors" table contained a record for each of the 798,673 registered users that have made an edit to an article or a discussion page. Furthermore, we collected the information on the number of days since they started contributing, the number of edits, the length of their contributions.

Figure 1, in the Annexes, summarizes our workflow. We detail in the following sub-sections the labeling of the articles and the characterization of the profiles of the authors of the GA/FA.

3.2 Dependent Variable : Unlabeled Articles Classification

After preparing the data, we built our training dataset by labeling our Article dataset. Our target variable, article quality, takes a value of *1* when the article is labeled as FA/GA and *0* otherwise. Since there is no information that supports whether an article is of poor quality if it is not labeled as FA/GA we considered them as unlabeled data. The number of articles with a classification represented less than 1% of the data. Therefore, we needed to obtain more classified data as a source of our final classifier. To carry out this task we obtained the article's length from the API and created a new attribute called "contrib_per_age" which is the division between "article_length_contribution" and "article_age".

6. http://dumps.wikimedia.your.org/eswiki/20151102/
7. https://github.com/glimmerphoenix/WikiDAT
8. https://es.wikipedia.org/wiki/Wikipedia:Estaditicas
9. https://www.mediawiki.org/wiki/API:Main_page

To classify our unlabeled data, we decided to use a One-Class SVM classifier [28], implemented in R [10]. We decided to exclude 10% of the FA/GA for validation purposes (as spies). To choose parameters, we looked for the number of False Negatives (the FA/GA classified as non FA/GA) and tried to decrease it. Finally, we trained with 3511 FA/GA and obtained from the unlabeled data : 238,810 labeled as positive and the rest (2,778,565) as negative. To validate our process, we included 497 of the Featured Articles and 453 of the Good Articles as unlabeled and then we validated whether they were correctly recognized.

We correctly classified 76% of the Featured Articles and 83% of the Good Articles. At this point, we had our 2,782,076 articles labeled. The execution of this algorithm with that quantity of articles did not take more than 2 minutes. This is less good than [39] as they had more variables, but mixing the explained variable and the explanatory variables. It is as good as similar classifications (i.e. article content variable based), such as [21], which identifies around 75% of FA with word count algorithms (but more than 90% when adding style variables, something we did not test here).

3.3 Explanatory Variables : Team Characteristics

With this labeled dataset we set up a feature engineering process over the contributors, to be able to take into account the number of FA/GA which the contributor had worked on. For this, we created the independent variables described above.

3.3.1 Efficiency

We measured the efficiency of the authors based on their contribution to good/feature articles. We computed the ratio between the number of FA/GA the contributor had worked on over the total number of Articles where the contributor had worked on.

$$efficiency_{contrib} = \frac{\#FA/GA\_worked\_in}{\#Articles\_worked\_in}$$

3.3.2 Contribution

We measured editor's contribution as the total number of edits performed by the editor on articles within the scope of the WikiProject.

3.3.3 Dispersion in participation

We measured editor's participation by the total number of concurrent projects of which the editor is currently listed as a member. A higher number means that the editor is involved in more projects at the same time.

3.3.4 Tenure

We measured editor's tenure as the amount of time a member had been active in Wikipedia (in days from the first edit).

3.3.5 Team level

In addition to calculating the variables for each contributor, we calculated them for the "Team", i.e. the group of contributors that worked on an article.

3.4 Final dataset

The Team table contained 20,717,227 rows, where each row was a pair article-contributor, denoting authorship. In other words, one single article appeared in as many rows as the number of authors it had.

10. http://www.inside-r.org/node/57517

Finally, we built our final dataset, 'Team Articles' in which we aggregated information about each contributor. The final structure of the variable used is presented in Table 1.

Table 1: Team Articles dataset

Variable Name	Attributes	Type	Description
F0 :	Min efficiency	Float	The min efficiency of its members
F1 :	Max efficiency	Float	The max efficiency of its members
F2 :	Mean efficiency	Float	The mean efficiency of its members
F3 :	Sum of contribution	Int	The sum of the number of contributions (edits) made by the team's members
F4 :	Total length of contribution	Int	The total length in terms of characters of the contributions made by the team-members
F5 :	Sum page revision	Int	Count of revision for different pages
F6 :	Sum participation	Int	Count of revision only for articles
F7 :	Min Tenure	Float	The experience of the least experienced contributor
F8 :	Max Tenure	Float	The experience of the most experienced contributor
F9 :	Mean Tenure	Float	The mean experience of the contributors
G1	Article quality	Int	1 : Featured Article & Good Articles, 0 : Non Featured Article

This derived dataset was highly unbalanced, making it difficult to run classifiers. So we ran the SMOTE algorithm [7], which aims at over-sampling the minority elements in a dataset classified as FA/GA. Doing so, we got the label's difference (regarding G1 variable) from 10 :90 to approximately 50 :50.

3.5 Data analysis and model tuning

After data preparation, we proceeded to the execution of a classification task using a Decision Tree to retrieve the rules and thus the most pertinent factors of the contributors to FA/GA. The idea of a decision tree is to classify by partitioning the input space into small segments and label these small segments with one of the various output categories.

To evaluate the performance of our Decision Tree, we measured over our validation dataset (spies) the False Negative Rate (FNR) that is the proportion of events that are being tested for which it yielded a negative outcome.

$$FNR = \frac{\#FA/GA\_bad\_classif.}{\#FA/GA\_good\_classif. + \#FA/GA\_bad\_classif.}$$

Using this measure of error, we tuned the model to obtain an 'optimal' solution, by changing the parameters inherent to the Decision Tree algorithm implemented on Spark. For instance, it was possible to modify the following ones : impurity measure (possible values being Gini and gain information [11]), max depth of the tree, and a maximum number of bins.

For the remaining parameters, we needed a tree that would not lose its readability and still provide good results ; therefore, we chose a maximum depth of 7, since increasing it did not improve the results significantly, and would have made it more difficult to extract the rules and see the results. Regarding the maximum number of bins, the default value is 32. As our variables were all continuous we could choose a lower number of bins and the one that yielded best results was 3.

4. RESULTS

Table 2, in the Annexes, presents the resulting rules from the decision tree that was tested over the FA/GA left for validation. The decision tree was composed of many rules from which only those that classified the articles as FA/GA were extracted. This extraction was done manually since the output of the decision tree was in plain text. Each of these is composed of an IF condition and, for each feature, a range of values. This means that, under this rule, an article which has its features between the values indicated by the rule is classified as FA/GA by the decision tree.

After split validation of the train set, we obtained a true positive rate of 91% and a true negative rate of 84% (thus our classifier performs best at classifying the FA/GA than the non FA/GA). Globally, this leads to a FNR (False Negative Rate) of 8% for our validation dataset. For the evaluation of our rules, we included two measurements :

— FA/GA ratio : number of articles that followed the rule (and were actually FA/GA according to Wikipedia) over the number of articles that followed the rule. This allowed us to measure in what degree this rule could truly separate the articles by their quality.

— Retrieval : Number of FA/GA retrieved by the rule, over the total number of FA/GA, according to Wikipedia. It measured the amount of quality articles that are retrieved by the rule with respect to all the quality articles.

These metrics were included to be able to extract information about the efficiency of our rules in detection and extraction of the FA/GA.

Regarding the rules, we can state that the "Maximum efficiency" ($F1$) and 'Sum of contributions' ($F3$) features are the most important features to discriminate the teams producing FA/GA from the others. Two rules (12 and 7) account for nearly 70% of the FA/GA. With the addition of rules number 9, 5 and 8, they were able to recognize 85% of all the FA/GA in the validation set. They took into account the "Team maximum efficiency" feature that ranges values over 0.36, which means that the percentage of maximum participation of members on FA/GA is 36%.

The rules show the segmentation leading to FA/GA articles. If $F1 < 36.6\%$, in other words, if there is not at least one contributor used to produce FA/GA, the chances for having a FA/GA are very low.

If $36.6\% < F1 < 55.6\%$ there is a good chance of finding FA/GA, if at the same time, $F3 > 1,325,311$ pages, and $F2 > 27.8\%$ and $F0 > 0.152$ (rule 7). This means that if there are not any very experienced contributors in FA/GA production, a team where people are experienced enough, in terms of total of contribution ($F3$), but also in terms of producing FA/GA (in mean, as the team participant's ratio of FA/GA over all articles they have contributed to exceed 27.8%, $F2$), and where there are no newcomers in the FA/GA production business (as the minimum efficiency ratio has to exceed 15.2%, $F0$) can succeed. However, having this type of team is not a guarantee of success, since, for the articles were produced by teams with this profile, only 27% are FA/GA.

Otherwise, and for the biggest 'type' of FA/GA (45.7%), not

11. The official documentation recommended to use Gini's impurity measure, which yielded the best results, what we did here.

only $F1 > 55.6\%$ and $F2 > 36.5\%$, but also $F4 > 2.143e9$ and $F3 > 1,325,311$ (rule 12). In this configuration, there is a very experienced contributor to FA/GA (F1), and the team is very experienced in mean in that matter too (F2), and has contributed a lot of edits (F3) and a lot of content (F4). Having this kind of very experienced team with a strong 'leader' seems to be the perfect combination as more than 76% of the articles produced by this team profile are FA/GA.

However, rule 9 stresses that, a less experienced team in terms of mean FA/GA participation can succeed ($F2 < 36.5\%$), as long as it has a strong 'leader' ($F1 > 55.6\%$), a good record of content contribution, as a team ($F4 > 2.143e9$), and no contributor with no experience of FA/GA ($F0 > 15.2\%$). This kind of team represents nearly 6% of the FA/GA and has more than 45% chance to be associated with a FA/GA.

5. DISCUSSION

Focusing on the team's characteristics only, and not on the article's characteristics, results in a successful classifier of high-quality articles, sustaining also what we've seen in related works using more complex metrics [32].

This analysis, done on the Spanish Wikipedia, seems to confirm a result already found in other Wikipedia (en-Wikipedia), and in open-source software : the existence of a "core member effect" where a small group of highly active core members is responsible for most of the (FA/GA) contribution to the project [8, 3], not only about the labeled articles (FA), which are those where the "administrators" and the people very involved in Wikipedia had their word to say, but, more generally, about all the articles of quality (the articles we clustered as FA/GA).

As expected, efficiency is the most important attribute that leads to FA/GA, especially the fact that the contributions of one of the editors include 36% FA/GA. For that matter, team efficiency is also very important (mean team member percentage of FA/GA over total articles contributed to has to be over 28.8%).

These results in Spanish Wikipedia confirm the investigation of [26] (cited by [30]), which showed that FA in German Wikipedia are not necessarily good because they were written by a huge number of people, but because some contributors feel personally 'committed' to the article.

In addition to participating in FA/GA production or to having a high efficiency, having a team of very productive editors, i.e. with a lot of contribution, especially in terms of length, seems to be key for producing an article of quality. In fact, the number of characters reflects the real amount of data added or deleted from the Wikipedia by an author. So as expected, past contribution like previous productivity affect team performance and then article quality. Somehow, this past productivity can be diffused among the team, but with less chance of success (rule 7).

Contradictory to our hypothesis that members who are involved in multiple projects in Wikipedia and would do less for each project, participating in other projects do not necessarily have a negative effect on article quality. Our findings show that participation in concurrent projects has just a slight impact on the quality of an article (rule 1).

This seems to be in contradiction with [29], who showed that at WikiProject level (Astronomy, Maths, etc.) a lot of small contributors matter more than some big contributors for the growth of the project. In addition to being on the en-Wikipedia, this study focused at the sub-project level, not at the article level, and is not concerned with the quality, but by the volume of edits (so the activity, not what remains as real knowledge). This is enough to explain the differences, but it would be very interesting to do our study at WikiProject level, as they did.

Evidence from Wikipedia collaboration has shown that, although old-timers have the experience and skills to contribute, their effort or motivation is generally lower than that of newcomers [41]. However, [25] showed that greater average membership tenure relates to project efficiency in a positive way. Our results do not show that experience, or tenure, in terms of the amount of time that an individual has been part of the Wikipedia community, has any effect on article quality. This is probably because, for the articles of quality, this effect is not strong enough to be in the top seven discriminatory variables, and is hidden by the importance of the experience in the classification (in other words, this may seem, but after the seventh stage where we stopped our classifier).

In addition to the theoretical contribution, our findings have several important implications for practice in the context of knowledge creation in online epistemic communities.

The results suggest that online teams with more efficient contributors are more likely to enjoy an enhanced capacity for better performance. This finding suggests that WikiProject managers may want to monitor the teams creating new articles, and inform those which are expected to fail, about what they lack (FA/GA expert(s), tenured contributors, for instance). Another way to do so could be to inform these key contributors about "under-staffed" projects which could use their expertise [12]. But they also stress the importance of participants' experience (tenure, FA/GA participation, the total number of characters contributed) for the efficiency of the team. So, in addition to the initiatives to attract and welcome the newcomers, the nesting initiatives [13] are as important to train the future key contributors. Here again, inviting beginners to take part in projects with a good team of experts may be key for the emergence of the future big contributors.

It would be interesting to confirm these results on other Wikipedia language projects and generalize the above managerial implications. Before doing so, the analysis of the data may be improved too. In other works, we used data extracted from the dump at the date of the dump creation. That means that some contributions may have occurred after article creation or cooperation. But it would have made more sense to calculate, for each article and each team, the contribution, tenure, etc., before the creation of the article, to propose a predictor of the efficiency of the team. This is, even more, computer-intensive, but should be feasible with the methodology we proposed in this article. This leads to the conclusion of this work.

6. CONCLUSION

In this study, we were interested in the prediction of the quality of a Spanish Wikipedia article by analyzing the characteristics of the team which produced it.

First, we faced the challenge of working with large data sets since the Spanish Wikipedia dump is around 24GB. We used different extraction and processing tools to be able to work around this problem (i.e. WikiDAT and Spark). Secondly, to be able to train our classifier we needed a more balanced dataset than the original one since only 1% of the data was classified positively and the rest was unlabeled. We used a PUlearning data mining technique to create our dataset of classified articles with a positive rate of 83%.

12. The Wikimedia Foundation's experience of automatically proposing articles existing in English to be translated into French to targeted French contributors, according to their taste, can be seen as the first step in this direction. For the detail of the study, see the discussion here : https://meta.wikimedia.org/wiki/Research_talk:Increasing_article_coverage, and the scientific analysis of the experiment here : [40].

After obtaining our dataset, we created the features that would represent a team. These features included the number of contributions in pages and articles, the length of the contributions made and we proposed a new metric which would measure the efficiency of a team, by calculating the number of FA/GA that each member has worked on over the total number of articles s/he has participated in. At this point we faced our third challenge, the fact that the data was highly unbalanced by a difference of $10 : 90$ in favor of those classified as non FA/GA We applied the SMOTE algorithm which led us to a final (and bigger) dataset having a $50 : 50$ proportion.

We modeled our problem using a Decision Tree which helped us in recognizing the most important team characteristics and their values to predict good quality articles with a positive rate of 92% in the validation set. After analyzing the importance of the features presented in the resulting rules, we can conclude that it supports our hypothesis on which we based the construction of these features. It is natural to think that a good article is written by people with a high percentage of good articles to their credit, which is described by Team mean efficiency. Furthermore, if it is written by a group of contributors that have included an extended amount of content. [14], but also the studies of group creativity and efficiency [36], stressed that collaborative projects are made possible by the aggregation of various motivations and level of involvement. This may be true, but, according to our results, the newcomers and the registered-but-less-involved contributors are not that important for achieving quality in the production (of Wikipedia articles).

In addition to the refinement in the calculation of the characteristics of the teams that we pointed out in the discussion, several extensions can be proposed in this article and its methodology. Regarding the characteristics of the article team, we created 7 attributes. Creating more variables/attributes may be beneficial to discover more team/member aspects that would lead to a quality Wikipedia article, even if the short number of variables and the high level of detection show that simple measures are quite enough to identify a good team.

We will also improve our classifier on the article quality detection side (explained variable), and [21]'s style variables are a natural direction on that matter.

Finally, looking not only at the characteristics of the team, but also on the process of production, and the role people take in that process, could also improve the detection of the teams of good quality, following the work of [11] [13].

As far as the analytic workflow process is concerned, we propose to implement One-Class SVM using scikit-learn [14] with the integration package for Spark [15] in order to use only one technology, since the transition from CSV, process, then back to CSV to transfer from R to Python was inefficient and prone to mistakes (and since our implementation over the E1071 R package [16] performed poorly). In this configuration, the pandas dataframes [17] and the numpy arrays [18] under-performed in comparison with the Spark resilient distributed dataset (RDD) [19].

7. REFERENCES

13. We thank one of the anonymous reviewers for this idea.
14. http://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html#sklearn.svm.OneClassSVM
15. https://databricks.com/blog/2016/02/08/auto-scaling-scikit-learn-with-spark.html
16. http://www.inside-r.org/node/57517
17. http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html
18. http://docs.scipy.org/doc/numpy-1.10.0/reference/generated/numpy.array.html
19. https://spark.apache.org/docs/latest/quick-start.html#basics

[1] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to wikipedia content. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym 2008, pages 26 :1–26 :12, New York, NY, USA, 2008. ACM.

[2] O. Arazy and O. Nov. Determinants of Wikipedia quality : The roles of global and local contribution inequality. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, pages 233–236, 2010.

[3] O. Arazy, O. Nov, R. Patterson, and L. Yeo. Information quality in Wikipedia : The effects of group composition and task conflict. *Journal of Management Information Systems*, 27(4) :71–98, 2011.

[4] A. B. Bakker, K. I. Van Der Zee, K. A. Lewig, and M. F. Dollard. The relationship between the big five personality factors and burnout : A study among volunteer counselors. *The Journal of social psychology*, 146(1) :31–50, 2006.

[5] J. E. Blumenstock. Size matters : Word count as a measure of quality on wikipedia. In *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, pages 1095–1096, 2008.

[6] L. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal Evolution of the Wikigraph. In *Proceedings of Web Intelligence*, pages 45–51, Hong Kong, Dec. 2006. IEEE CS Press.

[7] N. Chawala, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote : Synthetic minority over-sampleing technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

[8] K. Crowston and J. Howison. Hierarchy and centralization in free and open source software team communications. *Knowledge, Technology & Policy*, 18(4) :65–85, 2006.

[9] K. Crowston, J. Howison, and H. Annabi. Information system Success in Free and Open Source Software Development : Theory and Measures. *Software Process Improvement and Practice*, 11 :123–148, 2006.

[10] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3) :37, 1996.

[11] O. Ferschke, D. Yang, and C. Rosé. A lightly supervised approach to role identification in wikipedia talk page discussions. 2015.

[12] A. J. Flanagin and M. J. Metzger. From encyclopædia britannica to wikipedia : Generational differences in the perceived credibility of online encyclopedia information. *Information Communication and Society*, 14(3) :355–374, 2011.

[13] A. Forte, N. Kittur, V. Larco, H. Zhu, A. Bruckman, and R. E. Kraut. Coordination and beyond : social functions of groups in open content production. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 417–426, New York, NY, USA, 2012. ACM.

[14] C. Hess and E. Ostrom. Introduction : An Overview of the Knowledge Commons. In [15], editor, *Understanding Knowledge as a Commons. From Theory to Practice*, pages 3–26. 2006.

[15] C. Hess and E. Ostrom, editors. *Understanding Knowledge as a Commons. From Theory to Practice*. MIT Press, december 2006.

[16] N. Jullien, K. Crowston, and F. Ortega. The rise and fall of an online project : is bureaucracy killing efficiency in open knowledge production ? In *Proceedings of the 11th International Symposium on Open Collaboration, OpenSym*, page 13. ACM, 2015.

[17] W. A. Katz. *Introduction to reference work, vol. 1 : Basic Information Services*. McGraw-Hill, Boston, MA, 8th edition, 2002.

[18] D. Lewandowski and U. Spree. Ranking of Wikipedia articles in search engines revisited : Fair ranking for reasonable quality ? *Journal of the American Society for Information Science and Technology*, 62(1) :117–132, 2011.

[19] X. Li, J. Tang, T. Wang, Z. Luo, and M. de Rijke. Automatically assessing wikipedia article quality by exploiting article–editor networks. In *Advances in Information Retrieval*, pages 574–580. Springer, 2015.

[20] D. Lindsey. Evaluating quality control of wikipedia's feature articles. *First Monday*, 15(4), 2010.

[21] N. Lipka and B. Stein. Identifying featured articles in wikipedia : writing style matters. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1147–1148, New York, NY, USA, 2010. ACM.

[22] R. Lopes and L. Carriço. The impact of accessibility assessment in macro scale universal usability studies of the web. In *W4A'08 : Proceedings of the 2008 International Cross-Disciplinary Conference on Web Accessibility, W4A*, pages 5–14, 2008.

[23] G. Marwell and P. Oliver. *The Critical Mass in Collective Action : A Micro-Social Theory*. Cambridge University Press, Cambridge, 1993.

[24] J. Mathieu, M. T. Maynard, T. Rapp, and L. Gilson. Team effectiveness 1997-2007 : A review of recent advancements and a glimpse into the future. *Journal of management*, 34(3) :410–476, 2008.

[25] X. Qin, P. Cunningham, and M. Salter-Townshend. The influence of network structures of wikipedia discussion pages on the efficiency of wikiprojects. *Social Networks*, 43 :1–15, 2015.

[26] V. Rateike, J. Rösner, L. Denks, and C. Eberts. Die entwicklung exzellenter artikel in der deutschsprachigen wikipedia-keine qualität ohne küchenchef, 2007.

[27] F. Rullani and S. Haefliger. The periphery on stage : The intra-organizational dynamics in online communities of creation. *Research Policy*, 42(4) :941–953, 2013.

[28] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, J. C. Platt, et al. Support vector method for novelty detection. In *NIPS*, volume 12, pages 582–588. Citeseer, 1999.

[29] J. Solomon and R. Wash. Critical mass of what ? exploring community growth in wikiprojects. In *International AAAI Conference on Web and Social Media*, 2014.

[30] K. Stein and C. Hess. Does it matter who contributes - a study on featured articles in the german wikipedia. In *Hypertext 2007 : Proceedings of the Eighteenth ACM Conference on Hypertext and Hypermedia, HT'07*, pages 171–174, 2007.

[31] M. C. Sturman. Searching for the inverted u-shaped relationship between time and performance : Meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management*, 29(5) :609–640, 2003.

[32] B. Stvilia, A. Al-Faraj, and Y. Yi. Issues of cross-contextual information quality evaluation–The case of Arabic, English, and Korean Wikipedias. *Library & Information Science Research*, 31(4) :232–239, 2009.

[33] B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smith. A framework for information quality Assessment. *Journal of the American Society for Information Science and Technology*, 58(12) :1720–1733, 2007.

[34] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Information Quality Work Organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6) :983–1001, 2008.

[35] Y. Tan, V. Mookerjee, and P. Singh. Social capital, structural holes and team composition : Collaborative networks of the open source software community. *ICIS 2007 Proceedings*, page 155, 2007.

[36] B. Uzzi. A social network's changing statistical properties and the quality of human innovation. *Journal of Physics A : Mathematical and Theoretical*, 41(22) :224023, 12pgs, June 2008.

[37] F. B. Viégas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type : Coordination in Wikipedia. In *Proceedings of HICSS 2007*, pages 78–87, 2007.

[38] L. S. Wang, J. Chen, Y. Ren, and J. Riedl. Searching for the goldilocks zone : trade-offs in managing online volunteer groups. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 989–998, 2012.

[39] M. Warncke-Wang, D. Cosley, and J. Riedl. Tell Me More : An Actionable Quality Model for Wikipedia. *Proceedings of the 9th International Symposium on Open Collaboration, WikiSym '13*, pages 8 :1–8 :10, 2013.

[40] E. Wulczyn, R. West, L. Zia, and J. Leskovec. Growing Wikipedia Across Languages via Recommendation. *ArXiv*, Apr. 2016.

[41] W. Yang, W. Liu, A. Viña, M.-N. Tuanmu, G. He, T. Dietz, and J. Liu. Nonlinear effects of group size on collective action and resource outcomes. *Proceedings of the National Academy of Sciences*, 110(27) :10916–10921, 2013.

Table 2: Decision Tree Generated Rules

#	Rules	FA/GA ratio	Retrieval
1	If (F1 $\leqslant$ 0.366) AND (F4 >2.143e9) AND (F0 $\leqslant$ 0.176) AND (F0 $\leqslant$ 0.152) AND (F6 $\leqslant$ 879571.0) AND (F5 >966186.248) AND (F2 >0.278)	0.5	4.126e-06
2	If (F1 $\leqslant$ 0.366) AND (F4 >2.143e9) AND (F0 $\leqslant$ 0.176) AND (F0 >0.152) AND (F2 >0.278) AND (F3 >1325311.0) AND (F6 >2873122.0)	0.256	8.253e-05
3	If (F1 >0.366) AND (F1 $\leqslant$ 0.556) AND (F3 $\leqslant$ 1325311.0) AND (F4 >2.143e9) AND (F2 >0.278) AND (F2 $\leqslant$ 0.365) AND (F5 >966186.248)	0.114	0.00035
4	If (F1 >0.366) AND (F1 $\leqslant$ 0.556) AND (F3 $\leqslant$ 1325311.0) AND (F4 >2.143e9) AND (F2 >0.278) AND (F2 >0.365) AND (F0 $\leqslant$ 0.176)	0.18559	0.000627
5	If (F1 >0.366) AND (F1 $\leqslant$ 0.556) AND (F3 >1325311.0) AND (F2 >0.278) AND (F0 $\leqslant$ 0.152) AND (F2 $\leqslant$ 0.365) AND (F6 >879571.0)	0.11567	0.0550
6	If (F1 >0.366) AND (F1 $\leqslant$ 0.556) AND (F3 >1325311.0) AND (F2 >0.278) AND (F0 $\leqslant$ 0.152) AND (F2 >0.365)	0.2910	0.00645
7	If (F1 >0.366) AND (F1 $\leqslant$ 0.556) AND (F3 >1325311.0) AND (F2 >0.278) AND (F0 >0.152)	0.273	0.27412
8	If (F1 >0.556) AND (F4 >2.143e9) AND (F2 $\leqslant$ 0.365) AND (F0 $\leqslant$ 0.152) AND (F2 >0.278) AND (F6 >879571.0)	0.1746	0.0316
9	If (F1 >0.556) AND (F4 >2.143e9) AND (F2 $\leqslant$ 0.365) AND (F0 >0.152)	0.454	0.0585
10	If (F1 >0.556) AND (F4 >2.143e9) AND (F2 >0.365) AND (F3 $\leqslant$ 1325311.0) AND (F0 $\leqslant$ 0.152) AND (F5 $\leqslant$ 966186.248)	0.2798	0.000841
11	If (F1 >0.556) AND (F4 >2.143e9) AND (F2 >0.365) AND (F3 $\leqslant$ 1325311.0) AND (F0 >0.152)	0.3419	0.0135
12	If (F1 >0.556) AND (F4 >2.143e9) AND (F2 >0.365) AND (F3 >1325311.0)	0.76349	0.4575

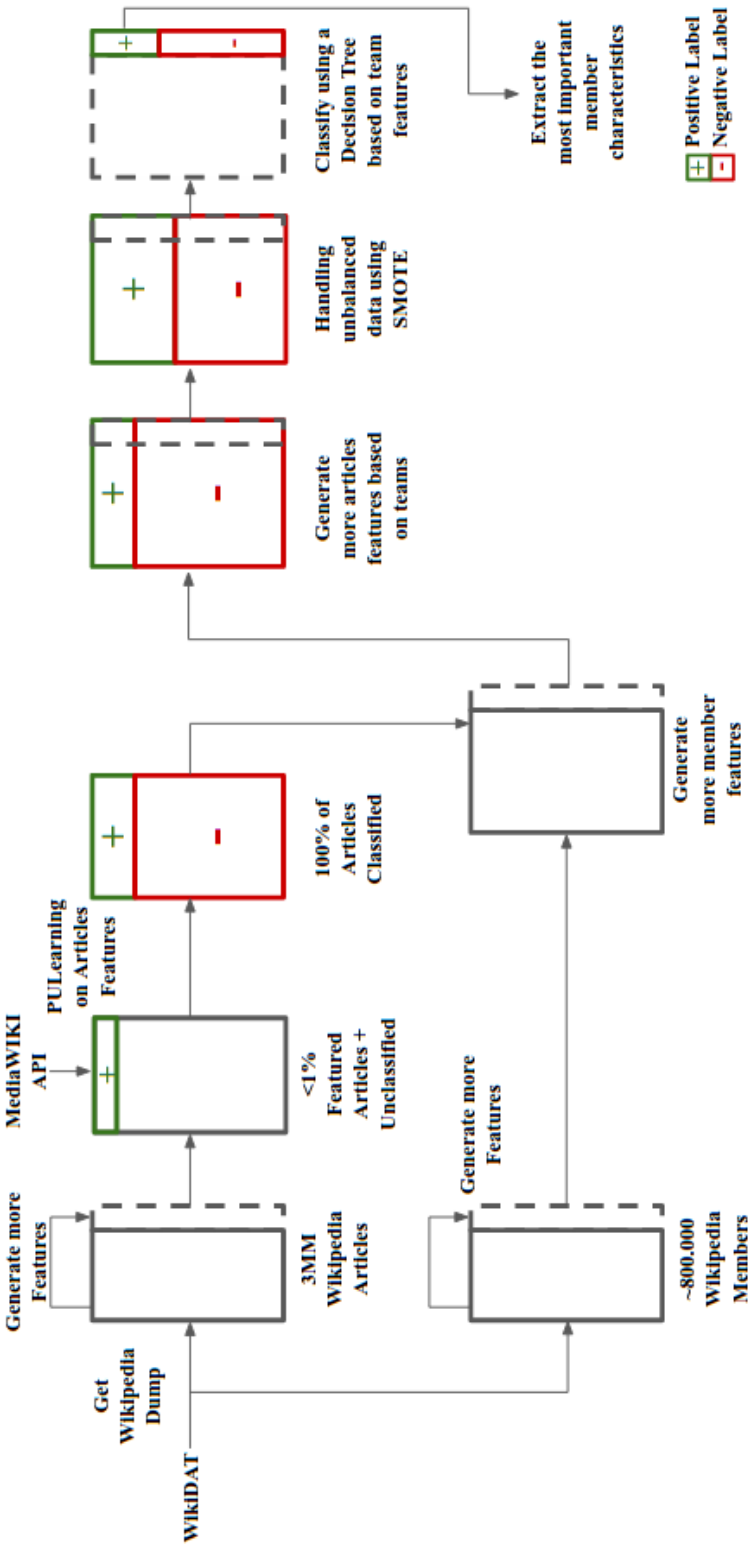

Figure 1: Data Analysis Process