Relative Quality Assessment of Wikipedia Articles in Different Languages Using Synthetic Measure *

Włodzimierz Lewoniewski, Krzysztof Węcel

Poznań University of Economics and Business, Poland
{wlodzimierz.lewoniewski,krzysztof.wecel}@ue.poznan.pl

Abstract. Online encyclopedia Wikipedia is one of the most popular sources of knowledge. It is often criticized for poor information quality. Articles can be created and edited even by anonymous users independently in almost 300 languages. Therefore, a difference in the information quality in various language versions on the same topic is observed. The Wikipedia community has created a system for assessing the quality of articles, which can be helpful in deciding which language version is more complete and correct. There are several issues: each Wikipedia language can use own grading scheme and there is usually a large number of unevaluated articles. In this paper, we propose to use a synthetic measure for automatic quality evaluation of the articles in different languages based on important features.

Keywords: Wikipedia, article quality, synthetic measure, wikirank.

1 Introduction

The social nature of Web 2.0 services offers almost all users the same freedom to contribute. Wikipedia one of the best examples of online collaborative human knowledge on the Web. This online encyclopedia has more than 44 million articles in almost 300 language editions.[1] English version is the biggest and have more than 5,4 million articles. There are other language versions, which consist over million articles, e.g. German, French, Russian, Polish.

There are systems of grades for article quality in Wikipedia and particular language version can use own assessment standard [1]. Each language version have special awards for articles with the best quality. In English version such articles are called "Featured articles" (FA). In German Wikipedia articles with the highest quality have name "Exzellente Artikel", what is essentially equivalent to FA grade in English. Such articles should be well written, in particular fulfil certain criteria. Articles that meet a core set of editorial standards but are not featured articles, qualify as "Good articles" (GA); in German language – "Lesenswerte Artikel". There also other lower quality grades. In English Wikipedia A-class, B-class, C-class, Start and Stub articles. However, quality grade scheme depends on language version. For example, German

[1] https://meta.wikimedia.org/wiki/List_of_Wikipedias

* This is a preprint version. The final publication is available at Springer via https://doi.org/10.1007/978-3-319-69023-0_24

Wikipedia not use other grades than FA and GA, Belarusian Wikipedia use only 3 grades (FA, GA, Stub).

Usually in each language version of Wikipedia there are only about 0,4-0,6% of high-quality articles (marked as FA or GA). Other articles can get lower quality grades but still most of the articles are unevaluated. For example, in Polish Wikipedia the share of articles without quality grade is about 99%. This number could be lowered by involving more experienced users and experts from different disciplines. Unfortunately, such experts are not always available.

Most of existing studies build quality models based on binary classification, which is limited in comparing articles on similar quality. In this work we propose to use synthetic measure to assess the quality of articles as continuous variable.

2 Related work

There are number of studies, which describes various ways to predict the quality of the Wikipedia articles. Some of them determine the quality based on article's content, another uses the edit history, the article's talk page and other sources. In general, we can divide related studies into the two groups: content-based and user-based approaches. Existing research works proposed different feature sets for measuring quality of Wikipedia articles.

Let's start by looking at scientific works analyzed the article content. One of the first studies showed that longer articles in Wikipedia often had higher quality grades [2]. Later works identified other features related to various constituents of the article: the best articles have more images, sections, use bigger number of references than articles with lower quality [1, 3, 4]. Special quality flaw templates can also help in articles assessment in Wikipedia [5].

In scientific works, attention is paid to writing style of articles, which depends on the language characteristics. High quality articles cover more concepts, objects and facts than weaker counterparts [6,7]. Thus, bigger relative number of facts in a document can indicate its higher informativeness. Character trigram feature can be used to analyze article writing style [8]. Another study used some basic lexical metrics derived from the statistic on word usages in Wikipedia articles as the factors that can reflect its quality [9]. Therefore, we can expect that high-quality articles use more nouns and verbs and less adjectives.

Other group of studies - works related to editor's behavior, explore how the users experience and coordinate their activities in relation to article quality. These approaches use various characteristics related to a user reputation and changes that they made [10,11]. Usually high quality articles have a large number of editors and edits [12]. Interaction among editors and articles can be visualized as a network, and using graph theory structural features associated to articles quality can be determined [13]. There is also artificial intelligence service involved to discover damaging edits, which can be used to immediately score the quality [14]. However, such user-based methods often require complex calculations and they do not analyze article itself, which would indicate what needs to be changed to improve its quality.

There are also a few works, that try to combine features from edition history and articles content [15, 16].

Concluding, existing studies propose different feature sets for assessing quality of articles in Wikipedia. However, there is no single universal feature set for doing it [16], especially if we consider different language versions [1,3]. It must also be taken into account that extraction and analysis of some features (e.g. lexical) depend on the language version [6,7,9].

We decided to consider only content-based features, because they can also show to Wikipedia contributors what can be changed in the article to improve it quality.

Majority of studies solve the problem of automatic quality assessment of articles as classification task: articles can be marked as Complete or Incomplete [1,3,4,6,7,9]. However, this approach is not able to show in what degree the article is better or worse than the other, if both are marked as the same class (e.g. Incomplete). An additional problem is caused by different standards in the quality grades between Wikipedia language editions.

Our work proposes to use synthetic measure to assess the articles' quality in different Wikipedia languages as a continuous variable. We verified our method on articles in 7 languages: Belarussian (BE), German (DE), English (EN), French (FR), Polish (PL), Russian (RU), Ukrainian (UK).

3 Building a synthetic measure

Proposed quality synthetic measure should be expressed as a real number between 0 and 100. So, the measure will cover the whole quality spectrum and relate quality to the highest quality class.

In order to build the synthetic measure we chose 5 important features, which were used in studies:

- Article length (in bytes)
- Number of references
- Number of images
- Headers 1st and 2nd level
- The ratio of number of references and article length.

As we mention before, in some Wikipedia language versions there are developed scale of grades. Often we can observe a positive correlation between the article quality and the value of each features. In English Wikipedia generally, the following quality classes are distinguished (from the highest): FA, GA, B, C, Start, Stub. Distribution of articles features of each quality class is shown in Figure 1. To build this chart we use randomly chosen 1000 articles from each quality class.

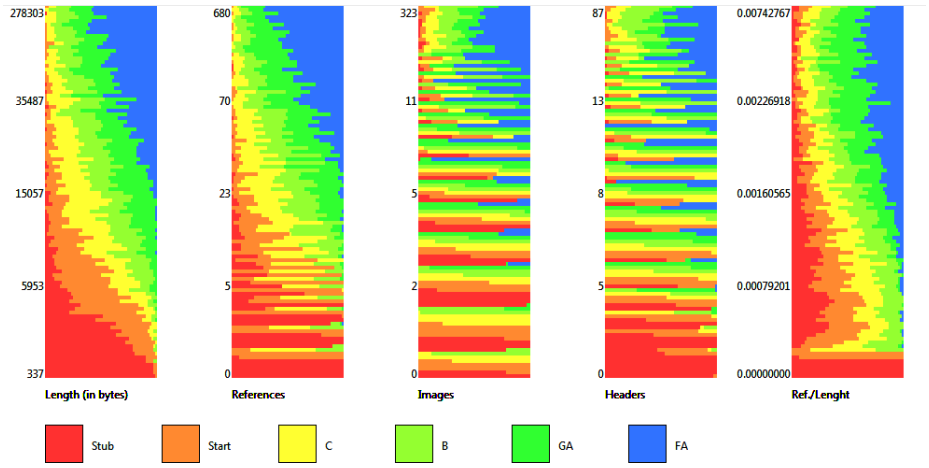

Figure 1. Distribution of features in articles of each quality class in English Wikipedia. Source: own calculation.

For any given feature and given language we calculate the median value in the highest quality class (FA). This value is used as a threshold. Medians for each considered feature and language versions are shown in Table 1.

Lang.	Length	References	Images	Headers	Ref./Len.
BE	198 365	210	36	27	0,001106
DE	56 238	55	17	21	0,000952
EN	49 038	115	13	14	0,002364
FR	91 004	185	29	26	0,002100
PL	59 672	96	17	17	0,001663
RU	139 415	163	24	22	0,001169
UK	82 371,5	40,5	24,5	21	0,000491

Table 1. Median feature values in the highest quality class in different Wikipedia languages. Source: own calculation.

Based on the presented medians we can normalize each feature in particular Wikipedia language version according to the following rule: if the value of the given feature in given language exceeds the threshold, it is set to 100 points, otherwise its value is linearly scaled to reflect the relation of the value to the median value. Let us assume that the median for the number of images in the highest class is 32. Any article with higher number of articles will score 100 for this feature; article with 16 images will get proportionally 50 points after normalizing.

We assume that all features can have the same effect on the value of our measure, therefore articles quality can be calculated according to the following formula:

$$Quality\ Score = \frac{1}{c}\sum_{i=1}^{c} nf_i$$

where:
nf_i – normalized feature *i,*
c – number of features.

So, she quality score is calculated as a average of single transformed variables, where weights are derived from significance of these variables as estimated by model. Example of using such synthetic measure can be observed in online service WikiRank[2], which use some of the content quantitative features to assess relative quality of Wikipedia articles in different languages.

4 Test Datasets

All the below mentioned datasets were created based on Wikipedia dumps from May, 2017. We use own parsers to extract particular features of articles.

4.1. LS

We decided to choose 3 datasets, which describe cities in selected countries: Poland, Germany, and France. Cities are usually best described in a mother tongue, therefore we call them language-sensitive (LS). For verification we choose cities, which are described in at least 5 languages: DE, EN, FR, PL, and RU. Therefore, we choose articles about 10516 German cities, 10092 French cities, 904 Polish cities.

In each LS dataset we count articles that have the highest particular feature and the highest quality score. Share of the best articles count is shown below.

German cities

	Length	References	Images	Headers	Ref./Len.	Score
DE	91,73%	96,87%	52,21%	80,88%	89,07%	95,49%
EN	7,00%	1,95%	18,70%	11,26%	0,78%	4,14%
FR	0,03%	0,56%	0,94%	0,03%	2,81%	0,26%
PL	0,79%	0,51%	11,11%	0,10%	4,72%	0,10%
RU	0,01%	0,06%	0,66%	0,01%	2,31%	0,02%

Table 2. Share of articles with the highest value of quality score and particular feature in various languages of Wikipedia in German cities dataset. Source: own calculation.

According to Table 2, more than 95% German cities are best described in German Wikipedia. If we consider individual features, it is noticeable that the images count is relatively the worst predictor among features of language affiliation of the selected articles – only about half of the articles that describe cities in Germany have the highest

[2] http://wikirank.net

number of images in their own language. Much better result shows number of references – that feature has even better prediction than quality score.

French cities

	Length	References	Images	Headers	Ref./Len.	Score
DE	12,50%	16,57%	0,37%	0,20%	23,47%	9,24%
EN	5,23%	4,11%	52,80%	3,77%	3,29%	5,44%
FR	73,50%	57,68%	39,32%	92,14%	36,84%	79,41%
PL	0,44%	0,93%	2,92%	0,00%	2,00%	0,28%
RU	7,95%	19,72%	0,15%	0,02%	30,84%	5,62%

Table 3. Share of articles with the highest value of quality score and particular feature in various languages of Wikipedia in French cities dataset. Source: own calculation.

Table 3 shows that almost 80% of French cities are the best described in their native language according to our synthetic measure. Similarly to the case of German cities, the number of images shows relatively low prediction – only less than 40% of articles that describe cities in France have the highest value of this feature in their own language. Moreover, over half of the articles in another language version (English) that describe French cities have the highest number of images value. It should be noted that in this dataset references to length ratio is the worst predictor in contrast to German cities dataset, where this feature shows over 80% prediction. Slightly less number of articles in another language versions (Russian) has also the one of the highest value of this feature. In French cities dataset, the number of headers has better predictive power than quality score– over 90% of articles about French cities have the highest value of that feature in their own language version. So, almost all articles of this dataset have larger number of sections in French Wikipedia, which may indicate a more comprehensive description of cities in comparison with other considered language versions.

Polish cities

	Length	References	Images	Headers	Ref./Len.	Score
DE	12,94%	31,42%	0,00%	5,86%	59,51%	13,94%
EN	1,33%	1,33%	0,22%	0,44%	3,21%	1,22%
FR	0,00%	0,00%	0,00%	0,22%	2,65%	0,00%
PL	83,41%	66,37%	70,13%	82,08%	29,98%	84,85%
RU	0,00%	0,11%	0,00%	0,00%	2,99%	0,00%

Table 4. Share of articles with the highest value of quality score and particular feature in various languages of Wikipedia in Polish cities dataset. Source: own calculation.

According to Table 4, the Polish dataset quality score has the highest prediction of language affiliation of considered articles than each individual feature. It is noticeable that according to this score almost 14% Polish cities are described better in German Wikipedia that in others languages. That can be explained by geographical location and relatively large popularity of some of these cities among German people. If we consider individual features with high precision ability, we can distinguish two of them: articles

length and headers count. Like quality score these features separately predict almost the same number of articles in the Polish version. However, in much more articles in German version of this dataset have the highest value of references to length ratio than in Polish language - the difference is about twice.

In LS datasets quality score calculated by proposed method shows high precision. Depending on topic individual parameters can also show even higher precision than synthetic measure. However, there is no universal parameter for all presented topics that solve this task. Therefore, synthetic measure use different features for quality assessment.

Now, let's try to assess the quality of articles that are presented in different language versions of Wikipedia and don't have distinct topic or language affiliation.

4.2. 5L Dataset

In this dataset we chose 273 878 articles, written in at least 5 languages: DE, EN, FR, PL, RU. According to Table 5 we see, that the largest number of the best quality articles is in English version - slightly more than half of the considered titles.

	Length	References	Images	Headers	Ref./Len.	Score
DE	19,59%	31,17%	9,56%	12,41%	38,81%	22,58%
EN	61,46%	41,70%	46,73%	57,95%	17,71%	53,34%
FR	7,33%	8,59%	16,38%	12,80%	10,04%	11,52%
PL	5,31%	6,69%	7,70%	5,26%	9,72%	6,36%
RU	4,40%	5,95%	9,11%	5,18%	10,74%	6,10%

Table 5. Share of articles with the highest value of quality score and particular feature in various languages of Wikipedia in 5L dataset. Source: own calculation.

English version also have the largest number of articles with the highest value of individual features except for the references to length ratio, which has highest value in almost 40% of German version. According to these indicators we can conclude, that the greater number of articles from English and German Wikipedia are more developed among 5 considered languages.

4.3. 7L Dataset

In this dataset we choose 46 957 articles, written in at least 7 languages: BE, DE, EN, FR, PL, RU, UK. From Table 6 we can confirm the findings of previous 5L dataset on the share of articles with highest values of features and quality score.

	Length	References	Images	Headers	Ref./Len.	Score
BE	0,10%	0,24%	0,10%	0,21%	2,11%	0,23%
DE	14,86%	20,38%	11,17%	7,59%	23,74%	17,15%
EN	57,32%	38,56%	43,55%	50,75%	14,87%	49,99%
FR	5,71%	10,29%	10,79%	13,18%	10,65%	9,77%
PL	4,21%	4,00%	4,63%	4,56%	6,10%	4,13%
RU	6,79%	5,06%	7,51%	5,51%	6,18%	6,08%
UK	4,36%	15,74%	2,65%	2,40%	19,15%	12,19%

Table 6. Share of articles with the highest value of quality score and particular feature in various languages of Wikipedia in 7L dataset. Source: own calculation.

Results from 7L and 5L dataset lead to general conclusion: English version of Wikipedia has the largest share of articles with the relatively better quality than other languages. German Wikipedia is in the second place by general relative quality of articles. This fact is also confirmed by other indicators of these language versions of Wikipedia – they have the largest quantity of edits and the greatest number of active users[3]. However, this rule does not apply to the Ukrainian Wikipedia, which has about 12% of articles with the highest quality score in 7L dataset despite the fact that this language version is less developed than French, Polish and Russian Wikipedia.

5 Articles Assessment

In this section we present the results of assessing over 10 million articles in 7 language versions based on Wikipedia dumps from May, 2017. Table 7 presents share of articles whose quality score falls within the specified interval.

Score interval	BE	DE	EN	FR	PL	RU	UK
[0,10)	72,29%	56,90%	30,35%	41,13%	49,88%	60,80%	59,31%
[10,20)	20,11%	7,65%	27,51%	30,25%	22,85%	20,30%	10,15%
[20,30)	5,95%	19,88%	22,23%	18,14%	18,17%	12,34%	16,96%
[30,40)	0,96%	10,10%	10,55%	6,92%	6,19%	4,43%	10,50%
[40,50)	0,36%	3,02%	4,50%	1,90%	1,68%	1,17%	1,66%
[50,60)	0,14%	1,19%	2,17%	0,80%	0,61%	0,47%	0,68%
[60,70)	0,07%	0,59%	1,18%	0,38%	0,30%	0,22%	0,32%
[70,80)	0,04%	0,30%	0,66%	0,20%	0,14%	0,11%	0,19%
[80,90)	0,04%	0,19%	0,45%	0,14%	0,09%	0,08%	0,13%
[90,100]	0,04%	0,18%	0,40%	0,13%	0,09%	0,07%	0,11%

Table 6. Share of Wikipedia articles whose quality score falls within the specified interval in each of seven language versions. Source: own calculation.

Results shows that in all language versions more than 90% of articles have quality score less than 40. The greatest number of articles that have quality score 40 and more is English and German Wikipedia.

More clear distribution of quality score is presented in Figure 2. We can see that articles whose quality score falls within the highest interval [90,100] usually have maximum value of synthetic measure.

[3] https://en.wikipedia.org/wiki/List_of_Wikipedias

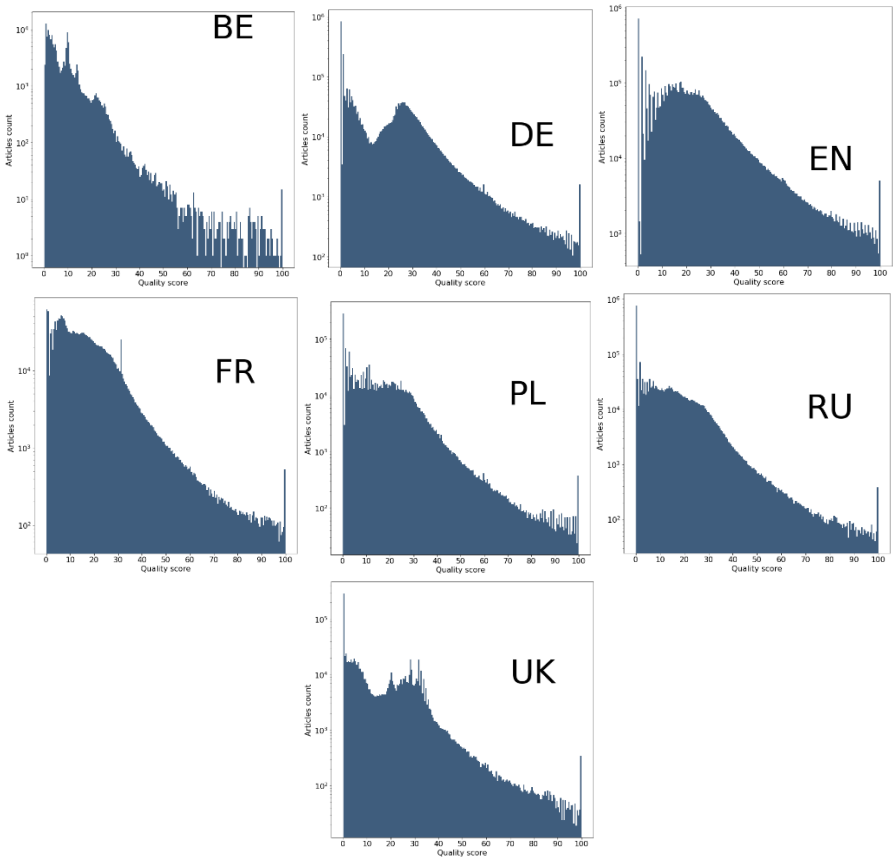

Figure 2. Distribution of articles with assessed quality score using synthetic measure. Source: own calculation.

6 Conclusions and Future Work

Synthetic measure can help to assess the quality of articles in different Wikipedia languages. In language-sensitive topics our approach can achieve precision over 90%. Differences between predicting ability of the individual features depending on topic shows that it is necessary to provide different weight for each component of the synthetic measure in each language version. In future we plan to extend the number of features and take into account their importance in particular language.

Quality assessment model can be applied in evaluation of the data quality placed in infoboxes.

One of the interesting directions of research is to examine the quality of information in relation to demand. It can be expected that the bigger the number of users reading a Wikipedia article, the bigger number of people interested in improving the content. So, the most popular language version can have also the best quality. Figure 3 presents example of comparison of popularity and quality of article about Kersti Kaljulaid in service WikiRank.

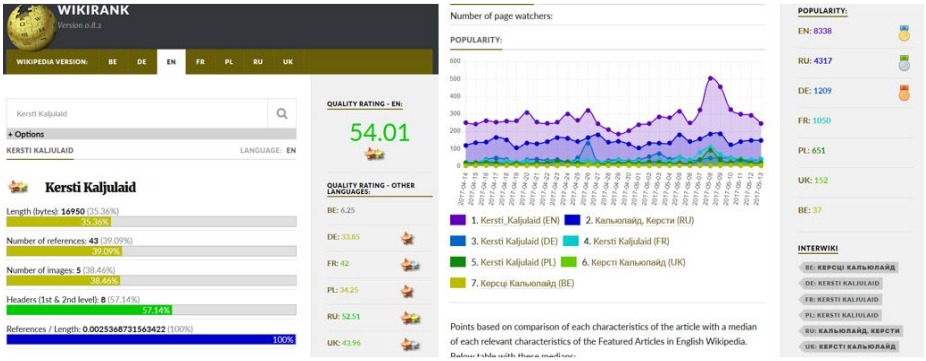

Figure 3. Quality and popularity comparison.
Source: http://wikirank.net/en/Kersti_Kaljulaid

References

1. Węcel, K., Lewoniewski, W., (2015), Modelling the Quality of Attributes in Wikipedia Infoboxes. In Business Information Systems Workshops. Volume 228 of Lecture Notes in Business Information Processing. Springer International Publishing, pp.308–320
2. Blumenstock, J., (2008), Size matters: word count as a measure of quality on wikipedia. In Proceedings of the 17th international conference on World Wide Web (pp. 1095-1096). ACM.
3. Lewoniewski,W., Węcel, K., Abramowicz,W., (2016), Quality and importance of Wikipedia articles in different languages. In: Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings. Springer International Publishing, Cham (2016) 613–624.
4. Warncke-Wang, M., Cosley, D., & Riedl, J. (2013, August). Tell me more: an actionable quality model for Wikipedia. In Proceedings of the 9th International Symposium on Open Collaboration (p. 8). ACM.
5. Anderka, M., (2013), Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Phd, Bauhaus-Universitaet Weimar Germany.
6. Lex E. et al. Measuring the quality of web content using factual information. Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality. – ACM, 2012. – C. 7-10.
7. Khairova N., Lewoniewski W., Węcel K., (2017), Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. In: Abramowicz W. (eds) Business Information Systems. BIS 2017. Lecture Notes in Business Information Processing, vol 288. Springer, Cham
8. Lipka, N., Stein, B., (2010), Identifying featured articles in wikipedia: writing style matters. In Proceedings of the 19th international conference on World wide web (pp. 1147-1148). ACM.
9. Xu Y., Luo T., (2011), Measuring article quality in Wikipedia: Lexical clue model. Web Society (SWS), 2011 3rd Symposium on. – IEEE – pp. 141-146.
10. Wu, G., Harrigan, M., Cunningham, P., (2011), Characterizing wikipedia pages using edit network motif profiles. In Proceedings of the 3rd international workshop on Search and mining user-generated contents (pp. 45-52). ACM.

11. Suzuki, Y., Nakamura, S., (2016), Assessing the Quality of Wikipedia Editors through Crowdsourcing. In Proceedings of the 25th International Conference Companion on World Wide Web (pp. 1001-1006). International World Wide Web Conferences Steering Committee.
12. Wilkinson, D. M., & Huberman, B. A, (2007), Cooperation and quality in wikipedia. In Proceedings of the 2007 international symposium on Wikis (pp. 157-164). ACM.
13. Ingawale, M., Dutta, A., Roy, R., Seetharaman, P. (2013). Network analysis of user generated content quality in Wikipedia. Online Information Review, 37(4), 602-619.
14. Halfaker, A., Taraborelli, D., (2015), Artificial intelligence service gives Wikipedians 'x-ray specs' to see through bad edits. https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs, Accessed: 25.04.2017
15. Dalip, D. H., Lima, H., Gonçalves, M. A., Cristo, M., Calado, P., (2014), Quality assessment of collaborative content with minimal information. In Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on (pp. 201-210). IEEE.
16. Dang, Q. V., Ignat, C. L., (2016), Quality assessment of wikipedia articles without feature engineering. In Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on (pp. 27-30). IEEE.