Quality and Importance of Wikipedia Articles in Different Languages

Włodzimierz Lewoniewski, Krzysztof Węcel, Witold Abramowicz

Poznań Univercity of Economics and Business,
Al. Niepodległości 10, 61-875 Poznań, Poland
wlodzimierz.lewoniewski@kie.ue.poznan.pl

Abstract. * This article aims to analyse the importance of the Wikipedia articles in different languages (English, French, Russian, Polish) and the impact of the importance on the quality of articles. Based on the analysis of literature and our own experience we collected measures related to articles, specifying various aspects of quality that will be used to build the models of articles importance. For each language version, the influential parameters are selected that may allow automatic assessment of the validity of the article. Links between articles in different languages offer opportunities in terms of comparison and verification of the quality of information provided by various Wikipedia communities. Therefore, the model can be used not only for a relative assessment of the content of the whole article, but also for a relative assessment of the quality of data contained in their structural parts, the so-called infoboxes.
Keywords: Wikipedia, DBpedia, information quality, data quality, WikiRank, article importance
JEL classification: C55, D8, L15, L86

1 Introduction

Currently there are 282 active Wikipedia language editions[1].The largest is the English version, which has more than 5 million articles. The first ten biggest editions also includes German, French, Russian and Polish.

This online encyclopedia has become one of the most important sources of knowledge throughout the world. In April 2016, the number of visits amounted to 282 million per day in all the language versions[2]. In the ranking of the most popular websites Wikipedia occupies 6th place in the world[3].

Every day increases the number of articles in each language. Articles can be created (edited) also by anonymous users. The authors do not have to formally demonstrate their skills in a specific field. Wikipedia has no central editorial or group of reviewers who could comprehensive approaches to verify all new and existing products. These

* This is a preprint version. The original publication available at http://dx.doi.org/10.1007/978-3-319-46254-7_50
[1] https://en.wikipedia.org/wiki/List_of_Wikipedias
[2] https://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm
[3] http://www.alexa.com/topsites

and other problems led to criticism of the concept of Wikipedia, in particular pointing out the poor quality of information[4].

Quality issues, however, concern the creators of Wikipedia. Practically every language version of the online encyclopedia have an award system for high quality articles. In the English version of Wikipedia the best articles have name „Featured Article" (FA). Articles that does not fulfill all the criteria FA, but closer to their quality, they can also get slightly lower award „Good Article" (GA).

In order to receive award article must be submitted for nomination by the user. The result of this is carried out discussion and voting takes place, where every user can approve or not to give awards for the specific article and explain their point of view. The criteria and rules for granting awards in each language version may change over time, which in turn may result in loss of award by some articles[5].

In addition to the award, in some language versions the article may receive lower scores. Such an indirect assessment may indicate,, maturity " of the article (i.e., in what degree it is close to the best articles). The English version of Wikipedia generally distinguishes 7 quality classes of articles (from the highest): FA, GA, A-class, B-class, C-class, Start, Stub. It is noteworthy that, unlike higher classes FA and GA, the other (lower) grades are received without a community discussion and voting – each user can set the rating by himself on the basis of rules. Some language versions use less-developed grading scale, e.g. in Polish version in addition awarded equivalent FA and GA are also grades[6]: Czwórka, Start, Zalążek (altogether 5 classes).

In Wikipedia there is no generally accepted standard classification of quality articles between different language versions [1]. Some languages use expanded rating scale (EN, RU), others are limited to 2-3 grades (BE, DE).In other words, each language version can have its own classification system of articles quality, but all of them use at least two highest classes - equivalent for FA and GA. However, such articles are very few - on average, in each language version of their share is about 0,07%. It should also be noted that a large part of the articles is not even evaluated, eg. in Polish edition share of such articles is over 99%.

In some language there is an importance scale[7] for articles. This feature is used for rating article importance in particular subject (or subjects) and usually marked as Top-, High-, Medium- or Low-importance. It can be expected that the greater the importance of the article, the better its quality. However, it should be taken into account quality class. Figures 1 and 2 show summary table by taking quality and importance rating for each assessed article in English, French, Russian and Polish Wikipedia. In contrast to similar statistics at Wikipedia[8], where one article could count to 2 times or more, we took into account only one of the highest quality and importance grade of each

[4] https://en.wikipedia.org/wiki/Criticism_of_Wikipedia

[5] For English Wikipedia there is a list of articles that have lost their award - https://en.wikipedia.org/wiki/Wikipedia:Former_featured_articles

[6] https://pl.wikipedia.org/wiki/Szablon:Stopnie_oceny_jako%C5%9Bci

[7] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wikipedia/Assessment

[8] https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team

article. On it to the reason, for example, the class „A" has only 181 articles, although technically number of such articles - 1593. It is connected with that the vast majority of articles with grade „A" are additionally evaluated as FA or GA. Therefore, in our experiments we will not take into account the class „A".

Quality	Importance (English Wikipedia)					
	Top	High	Mid	Low	???	Total
FA	854	1 363	1 450	916	185	**4 768**
GA	1 543	3 790	8 219	8 996	1 702	**24 250**
A	12	35	88	36	10	**181**
B	7 872	16 738	28 602	25 524	11 278	**90 014**
C	8 927	25 511	56 008	75 987	37 855	**204 288**
Start	13 117	61 840	259 776	654 384	254 221	**1 243 338**
Stub	3 027	24 139	190 874	1 594 570	706 921	**2 519 531**
???	2 130	8 770	28 198	84 995	945 621	**1 069 714**
Total	**37 482**	**142 186**	**573 215**	**2 445 408**	**1 951 793**	**5 156 084**

Quality	Importance (French Wikipedia)					
	Top	High	Mid	Low	???	Total
FA (AdQ)	410	494	396	211	-	**1 511**
GA (BA)	382	633	874	654	-	**2 543**
A	316	399	522	367	6	**1 610**
B	3 567	6 647	10 148	8 723	221	**29 306**
Start (BD)	6 355	24 405	71 048	139 354	6 657	**247 819**
Stub (É)	3 112	19 706	123 515	550 221	91 402	**787 956**
???	8	70	521	4 249	682 259	**687 107**
Total	**14 150**	**52 354**	**207 024**	**703 779**	**780 545**	**1 757 852**

Fig. 1. Articles by quality and importance in English (on the left) and French (on the right) Wikipedia. Source: own calculations in May 2016

Quality	Importance (Russian Wikipedia)					
	Top	High	Mid	Low	???	Total
FA (ИС)	79	111	140	13	549	**892**
GA (ХС)	82	224	370	76	1 659	**2 411**
НqA (ДС)	9	95	446	120	2 304	**2 974**
I	278	734	649	213	249	**2 123**
II	621	1 827	4 246	1 285	796	**8 775**
III	1 144	6 435	19 240	12 887	3 435	**43 141**
IV	698	4 165	27 717	17 262	8 872	**58 714**
???	688	3 039	5 213	6 999	1 176 737	**1 192 676**
Total	**3 599**	**16 630**	**58 021**	**38 855**	**1 194 601**	**1 311 706**

Quality	Importance (Polish Wikipedia)					
	Top	High	Mid	Low	???	Total
FA (AMN)	12	36	19	10	607	**684**
GA (DA)	18	37	36	27	2 008	**2 126**
B (Czw)	9	27	62	13	49	**160**
C (Popr)	9	38	53	15	449	**564**
Start (Dost)	97	180	518	258	204	**1 257**
Stub	40	83	323	524	683	**1 653**
???	304	1 459	1 812	3 159	1 155 318	**1 162 052**
Total	**489**	**1 860**	**2 823**	**4 006**	**1 159 318**	**1 168 496**

Fig. 2. Articles by quality and importance in Russian (on the left) and Polish (on the right) Wikipedia. Source: own calculations in May 2016

In the scientific literature we can find studies, which offer different approaches to the automatic evaluation of the quality of Wikipedia articles. Based on the different characteristics of highly-rated (awarded) articles it is possible to evaluate other. Text length, number of references, the number of images and other articles' features can help in the quality assessment.

The aim of our research is to answer to the following questions:

– Does article importance affect its quality?
– What parameters can help to assess the importance of the article automatically?
– Is there a difference between importance models in different languages?

Most of the research on models for the quality of Wikipedia articles is focused on the "largest" language – English. In this paper we consider 4 popular languages:

English (en), French (fr), Polish (pl), Russian (ru), which have introduced the templates for specifying article's importance. This allows us to build models that will be able to compare articles quality in different languages. Besides, this is the first study in which we have build importance models of the article and we will conduct a comparative analysis of these models in a different languages.

2 Automatic quality assessment

Since founding and with the increasing popularity of Wikipedia there are more and more scientific publications on the quality of the information. One of the first studies showed that the measurement of the volume of content can help determine the degree of maturity of the article [2]. Work in this direction show that generally higher quality articles are longer [3], use references in a coherent way, are edited by hundreds of editors and have thousands of editions [4, 5].

In addition to quantitative analysis, later research has focused on the qualitative analysis around the content of the article. In one of the works has been used so-called. FOG index readability, which determines the degree accessibility of text [6]. In cases where the volume of contents in articles is similar, better article will have more factual information [7]. Style and variety of words used also affects the quality of the article [8, 9]. Wikipedia users can include special templates in an article, indicating gaps in quality. Such annotations can help in assessing the quality of the article [10]. Features related to articles popularity can also be used in assessing the quality of the information they contain [11].

Another works on automatic quality classification of Wikipedia articles taking into account user behavior. There are models that take into account their experience and reputation. Articles quality has a large number of editing and a large number of editors who have a high level of cooperation [12, 13]. It is important that in this group of editors was even one user with a high level of experience in content editing in Wikipedia [14]. Particular importance have the reputation of the user who made the first edition of the article [15]. Reputation can be calculated on the basis of „survival" of the text, which user placed [16, 17, 18].

In this study, we decided to focus primarily on those aspects that can help improve the quality of the article – so we consider the content of the article and its metadata.

3 Data selection and extraction

On the basis of literature [19, 2, 4, 12, 3, 20, 6, 8, 21, 10, 22, 11] and our own research we have chosen 85 articles parameters which will be taken into account when building quality and importance models of Wikipedia articles. These parameters include various areas such as text statistics, parts of speech, readability formulas, similarity of words, the structure of the article, edition history, network parameters, popularity of the article, the characteristics of discussion.

One of the most attractive methods for obtaining data from Wikipedia is API service, which provides easy access to data and metadata of articles using HTTP, via a URL

in a variety formats (including XML, JSON). API service works for every language and is available at the address specified by the template: `https://{lang}.wikipedia.org/w/api.php?action={settings}`, where `{lang}` – abbreviation of the language version, `{settings}` – query settings[9]. Possibilities of API used in our specially prepared program WikiAnalyzer, which can get over 50 different parameters of each article.

In figure 3 the distribution of variables in articles with different quality class is shown. It is noticeable that the increase in each feature attracts increase in a share of higher-quality articles. We also compared parameters of the articles only from FA-class but of varying importance. Some of them are shown in figure 4 in English Wikipedia. Here regularity is also observed: the increase in value of features involves increases in a share of important articles.

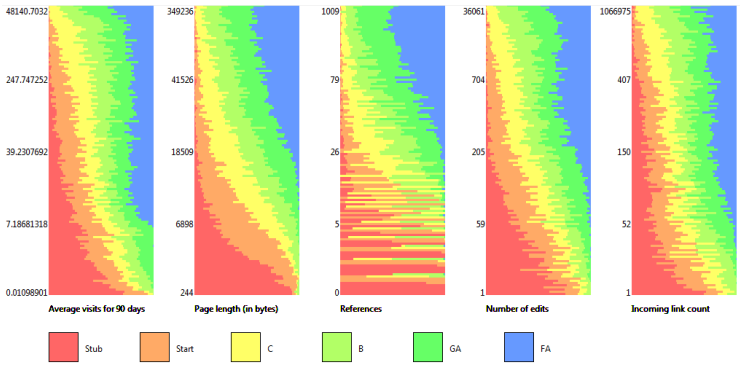

Fig. 3. Distribution of variables in articles with different quality class in English Wikipedia

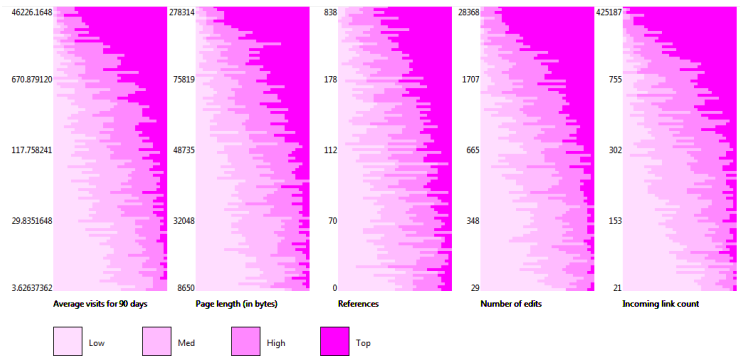

Fig. 4. Distribution of variables in FA articles with different importance in English Wikipedia

[9] All possible settings in API service can be found on a special page: `https://en.wikipedia.org/wiki/Special:ApiSandbox`

3.1 Dataset ENQ

For the answer to the first question raised in the introduction, we decided to make evaluation on articles with certain quality and certain importance of English Wikipedia, because this version:

- is the largest language version
- has the developed system of quality classification of articles
- has the greatest number of articles on intersections of quality and importance (see figure 1).

Because the smallest number of articles on intersection of quality and importance is 854 (for class FA and Top-importance) we decided to choose randomly 800 articles of each intersection (without A-class for the reasons described earlier). Altogether, there were 19200 articles in our ENQ dataset.

3.2 Datasets IMP

For the answer to the second and the third question evaluation in articles from different quality with certain importance. From 4 studied language versions the least developed system of importance assessment have Polish Wikipedia. There, the smallest number of Top-important articles was 489. Therefore, we have decided to choose randomly 400 articles from each importance level in each language version to allow the homogeneous distribution of the learning datasets.

4 Evaluation

In many approaches for building models the binary dependent variable was used [9, 7, 22, 11] and the quality was modelled as the probability of belonging to one of the two categories:

- **Complete** articles: FA-class and GA-class
- **Incomplete** articles: all other – developing (which should be further developed) and the unassessed articles.

Our previous research has shown that with such binary forecast variable the precision of 98-100% can be achieved (depending on the language version) [23]. Therefore, we decided to expand the number of alternatives in dependent variable – now each quality class is a separate name of this variable. For example, for our dataset ENQ we have 6 alternatives in dependent variable: FA, GA, B, C, Start, Stub.

Our researches have shown efficiency of Random Forest classifier on similar tasks, therefore in this study we also we use that data mining algorithm with default settings (100 trees, cross-validation with 10 folds) using WEKA software [23].

So, using 85 different articles parameters as independent variable and quality class as dependent we can reach 60% precision of classification. After inclusion of additional feature – importance of article – the precision of the model increased to 61%. Therefore

we conclude that inclusion of addition input variable (article importance) can improve the precision of classification.

The confusion matrix for model with quality class and importance level as dependent variables in ENQ dataset are shown in figure 5. Table 1 and 2 show the performance of the classifier. It can be argued that **importance of an article affects its quality**.

Observed quality	Predicted quality					
	FA	GA	B	C	Start	Stub
FA	**2 859**	277	52	11	1	0
GA	575	**2 302**	207	92	24	0
B	111	417	**1 261**	853	454	104
C	35	262	856	**1251**	699	97
Start	8	81	246	609	**1 734**	522
Stub	1	12	37	97	563	**2 490**

Observed importance	Predicted importance			
	Top	High	Mid	Low
Top	**3 176**	900	461	263
High	1 431	**1 608**	948	813
Mid	618	1 064	**1 559**	1559
Low	225	507	978	**3090**

Fig. 5. Confusion matrix - Quality (on the left) and Importance (on the right). English Wikipedia

Table 1. Classification results per quality class in English Wikipedia using Random Forest. Source: own study

Quality class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
FA	0.893	0.046	0.797	0.893	0.842	0.983
GA	0.719	0.066	0.687	0.719	0.703	0.946
B	0.394	0.087	0.474	0.394	0.43	0.827
C	0.391	0.104	0.429	0.391	0.409	0.827
Start	0.542	0.109	0.499	0.542	0.52	0.859
Stub	0.778	0.045	0.775	0.778	0.777	0.964
Overall	0.62	0.076	0.61	0.62	0.613	0.901

Now let's try to answer remaining two research questions. We use our IMP datasets which contains importance level as dependent variable. Using Random Forest as prediction model, we can obtain the most influential features, which affect article importance in each language. In figure 6 we show influence of each article parameter in importance model in different language editions of Wikipedia (in scale from 0 to 100, 100 - is the highest influence). As we can see, we have some differences between the models in particular languages. For example for English version the most influential features are: the sum of visits in 30 days, the number of links to article.

Table 2. Classification results per importance level in English Wikipedia using Random Forest. Source: own study

Importance level	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Top	0.662	0.158	0.583	0.662	0.62	0.852
High	0.335	0.172	0.394	0.335	0.362	0.676
Mid	0.325	0.166	0.395	0.325	0.357	0.672
Low	0.644	0.183	0.54	0.644	0.587	0.827
Overall	0.491	0.17	0.478	0.491	0.481	0.757

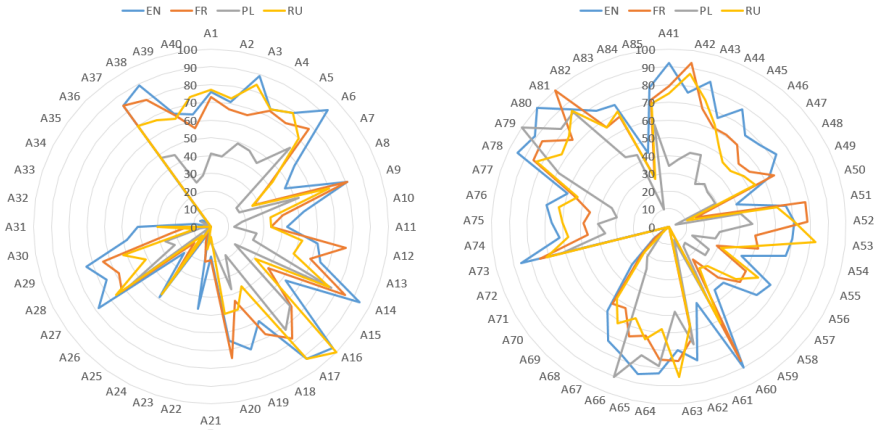

Fig. 6. Influence of article parameters in importance model in different laguage editions of Wikipedia (description of parameters abbreviations in table 3). Source: own study.

5 Conclusions

In this paper we have shown that the importance of the article affects the quality of the information contained in it. In our study we used ca. 80 features of articles and various data mining techniques to come up with a proposal for a quality models. We have also built the importance models for particular language edition of Wikipedia and shown the differences between these models.

The proposed models can help to improve the quality of Wikipedia articles by identifying the best version of a particular article. In consequence, our work can improve the quality of data in DBpedia[10], one of the most famous semantic database, which is enriched by extracting facts from articles of different language versions of Wikipedia. Data mining algorithms allow to determine the significance of the features in models of quality that can later be used to compare articles in different languages. This property is

[10] http://dbpedia.org

Table 3. Description of parameters abbreviations used in figure 6

Name	Description	Name	Description
A1	Last modified	A44	The number of pictures (all)
A2	Last modified not by the bot	A45	The number of unique pictures 1 lvl
A3	page length (in bytes)	A46	The number of unique pictures 2 lvl
A4	informativeness 1	A47	The number of unique pictures 3 lvl
A5	informativeness 2	A48	The number of unique pictures 4 lvl
A6	Number of edits by anonymous authors for the whole time	A49	The number of unique pictures 5 lvl
A7	Number of edits by anonymous for 12 months	A50	The number of followers
A8	Number of edits by anonymous for 6 months	A51	Number of templates (all)
A9	Number of edits by bots	A52	Number of templates ns10
A10	Number of edits by bots for 12 months	A53	Number of templates ns828
A11	Number of edits by bots for 6 months	A54	The number of unique anonymous for 12 months
A12	Number of edits for 12 months	A55	The number of unique anonymous for 6 months
A13	Number of edits for 6 months	A56	Number of unique authors for 12 months
A14	Number of edits for all time	A57	Number of unique authors for 6 months
A15	Number of edits for all time	A58	Number of unique bots for 12 months
A16	the number of links to the article (all)	A59	Number of unique bots for 6 months
A17	the number of links on the article ns0	A60	Number of unique bots for the all time
A18	the number of links on the article ns1	A61	Unique templates quality gaps
A19	the number of links on the article NS10	A62	Number of uhunique anonymous authors for the all time
A20	the number of links on the article NS100	A63	The number of language versions
A21	the number of links on the article ns101	A64	Median of non-zero last 30 days
A22	the number of links on the article ns11	A65	The median of visits for 30 days
A23	the number of links on the article NS12	A66	The median of visits for 90 days
A24	the number of links on the article ns13	A67	Heading 1
A25	the number of links on the article ns14	A68	Heading 2
A26	the number of links on the article NS15	A69	Heading 3
A27	the number of links on the article ns2	A70	Heading 4
A28	the number of links on the article ns3	A71	Heading 5
A29	the number of links on the article ns4	A72	Heading 6
A30	the number of links on the article NS5	A73	Come visit the ost day
A31	the number of links on the article NS6	A74	Ref / Length
A32	the number of links on the article ns7	A75	Ref / Number of letters
A33	the number of links on the article ns8	A76	References unique
A34	the number of links on the article ns828	A77	all references
A35	the number of links on the article ns829	A78	Average visits for 30 days
A36	the number of links on the article ns9	A79	Average visits for 90 days
A37	the number of internal links (all)	A80	Total visits for 30 days
A38	the number of good internal links	A81	Total visits for 90 days
A39	the number of broken internal links	A82	Noise1
A40	the number of external links	A83	Noise2
A41	The number of letters	A84	Unique authors for last time
A42	The number of letters without noise 1	A85	Unique authors for all time.
A43	The number of letters without noise 2		

used as the design creation service Wikirank[11], which is used to calculate the so-called relative quality of articles.

References

1. Węcel, K., Lewoniewski, W.: Modelling the Quality of Attributes in Wikipedia Infoboxes. In Abramowicz, W., ed.: Business Information Systems Workshops. Volume 228 of Lecture Notes in Business Information Processing. Springer International Publishing (2015) 308–320
2. Stvilia, B., Twidale, M.B., Smith, L.C., Gasser, L.: Assessing information quality of a community-based encyclopedia. Proc. ICIQ (2005) 442–454
3. Blumenstock, J.E.: Size matters: word count as a measure of quality on wikipedia. In: WWW. (2008) 1095–1096
4. Hu, M., Lim, E.P., Sun, A., Lauw, H.W., Vuong, B.Q.: Measuring article quality in wikipedia. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management - CIKM '07. (2007) 243–252
5. Wöhner, T., Peters, R.: Assessing the quality of Wikipedia articles with lifecycle based metrics. Proceedings of the 5th International Symposium on Wikis and Open Collaboration WikiSym 09 (2009) 1
6. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries. (2009) 295–304
7. Lex, E., Voelske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., Stein, B., Granitzer, M.: Measuring the quality of web content using factual information. Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality '12 (2012) 7
8. Lipka, N., Stein, B.: Identifying Featured Articles in Wikipedia: Writing Style Matters. Proceedings of the 19th International Conference on World Wide Web (2010) (2010) 1147–1148
9. Xu, Y., Luo, T.: Measuring article quality in Wikipedia: Lexical clue model. IEEE Symposium on Web Society (19) (2011) 141–146
10. Anderka, M.: Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. Phd, Bauhaus-Universitaet Weimar Germany (2013)
11. Lewoniewski, W., Węcel, K., Abramowicz, W.: Analiza porównawcza modeli jakości informacji w narodowych wersjach Wikipedii. In Porębska-Miąc, T., ed.: Systemy Wspomagania Organizacji SWO 2015. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach (2015) 133–154
12. Wilkinson, D.M., Huberman, B.a.: Cooperation and quality in wikipedia. Proceedings of the 2007 international symposium on Wikis WikiSym 07 (2007) 157–164
13. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in wikipedia. Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08 (2008) 37
14. Arazy, O.: Determinants of Wikipedia Quality : the Roles of Global and Local Contribution Inequality. New York (2010) 233–236
15. Stein, K., Hess, C.: Does it matter who contributes: a study on featured articles in the german wikipedia. HT '07: Proceedings of the eighteenth conference on Hypertext and hypermedia (2007) 171–174

[11] http://wikirank.net

16. Suzuki, Y., Yoshikawa, M.: Mutual Evaluation of Editors and Texts for Assessing Quality of Wikipedia Articles. In: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration. WikiSym '12, New York, NY, USA, ACM (2012) 18:1–18:10
17. Halfaker, A., Kraut, R., Riedl, J.: A Jury of Your Peers : Quality, Experience and Ownership in Wikipedia. WikiSym'09 (2009) 1–10
18. Adler, B.T., De Alfaro, L.: A content-driven reputation system for the wikipedia. Proceedings of the 16th international conference on World Wide Web WWW 07 **7**(Generic) (2007) 261
19. Lih, A.: Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. 5th International Symposium on Online Journalism (2004) 31
20. Blumenstock, J.E.: Automatically Assessing the Quality of Wikipedia Articles. Technical report (2008)
21. Dalip, D.H., Gonçalves, M.A., Cristo, M., Calado, P.: Automatic Assessment of Document Quality in Web Collaborative Digital Libraries. Journal of Data and Information Quality **2**(3) (2011) 1–30
22. Warncke-wang, M., Cosley, D., Riedl, J.: Tell Me More : An Actionable Quality Model for Wikipedia. In: WikiSym 2013. (2013) 1–10
23. Lewoniewski, W., Węcel, K., Abramowicz, W.: Analiza porównawcza modeli klasyfikacyjnych w kontekście oceny jakości artykułów wikipedii. In: VI Ogólnopolska Konferencja Naukowa. Matematyka i informatyka na usługach ekonomii im. Profesora Zbigniewa Czerwińskiego. (in press) (2016)