KDIR 2012

Proceedings of the
International Conference on
Knowledge Discovery and Information Retrieval

Barcelona, Spain

4 - 7 October, 2012

Sponsored by
INSTICC – Institute for Systems and Technologies of Information, Control and Communication

In Cooperation with
AAAI – Association for the Advancement of Artificial Intelligence

Edited by Ana Fred and Joaquim Filipe

Printed in Portugal

ISBN: 978-989-8565-29-7

Depósito Legal: 348249/12

http://www.kdir.ic3k.org/

kdir.secretariat@insticc.org

SPECIAL SESSION ON TEXT MINING

Full Papers

Short Papers

SPECIAL SESSION ON INFORMATION FILTERING AND RETRIEVAL: NOVEL DISTRIBUTED SYSTEMS AND APPLICATIONS

Full Papers

Visitors and Contributors in Wikipedia

Antonio J. Reinoso[1], Juan Ortega-Valiente[1], Rocío Muñoz-Mansilla[2] and Gabriel Pastor[1]

[1]*Department of ICT Engineering, UAX, Avda. de la Univerisdad, 1, Vva. de la Cañada, Spain*

[2]*Department of Computer Science and Automation, UNED, C/ Juan del Rosal, 16, Madrid, Spain*

{areinpei, jvalient}@myuax.com, rmunoz@dia.uned.es, gmartpas@myuax.com

Keywords: Wikipedia, Use Patterns, Traffic Characterization.

Abstract: Wikipedia continues to provide the community with a vast collection of articles that cover almost all the different areas of knowledge. The on-line encyclopedia is built upon altruistic contributions from individuals and organizations, which constitutes an absolutely new approach in knowledge compilation and distribution. The progression of Wikipedia to a mass phenomenon has promoted several initiatives devoted to deal with different issues concerning it, specially the quality of its contents and the authoring of its contributions. However, very few attention has been paid to aspects related to users' attitudes and behavior when browsing the information it offers. For this reason, this paper aims to find patterns that can describe how users interact and behave when visiting the Wikipedia's pages. We will consider the two most common forms of interaction between Wikipedia and its users: visits and contributions (edits). From these observations we will be used obtain different metrics, such as the degree of participation or the reluctance exhibited by users which, in addition, will be used to perform a comparison amongst the different Wikipedia editions. Our study is based on a sample of the requests that users submit to Wikipedia, which we receive in the form of log lines. Its results can help to better understand the nature of the relationship between Wikipedia and its users, and to properly characterize the different interactions between them.

1 INTRODUCTION

Wikipedia can be thought as a completely revolutionary approach for gathering and distributing knowledge. Its backing philosophy promotes a massive contribution and collaboration, as well as to join efforts in the process leading to the construction of any kind of knowledge. The resulting compendium of contents will remain available for the whole community, which will take benefit from it. The enormous interest attracted by Wikipedia can be appreciated from the non-stopping growth of its contents and from the huge number of visits that situates its website within the six most visited ones all over the Internet[1].

As a result of such popularity, Wikipedia has turned into a subject of interest for researchers[2]. However, most of this research is mainly focused on the reliability and quality aspects regarding the information offered by the Encyclopedia and on its growth and evolutionary tendencies. Our work, in contrast, aims to address the use given to Wikipedia by some of its most notorious communities of users through the analysis of the two most common forms of interactions demanded by users: visits and contributions.

Therefore, we have analyzed in this paper several issues related to the use given to different editions of Wikipedia by their corresponding communities of users. In particular, we will examine users' behavioral habits that can be extracted from the requests they submit when browsing Wikipedia. These habits include both general attitudes, like participation or collaboration, as well as more particular conducts, such as the realization of previewing of changes when editing articles or the issue of search operations from the Wikipedias's own search engine. As it is clear that different Wikipedias may present very different behavioral patterns in the interaction that they maintain with their respective users, we will compare the results obtained for different editions to analyze the differences and similarities among them.

Our results aim to present the observed patterns related to users' interactions with the on-line encyclopedia and their most common attitudes towards it. In particular, contributions (edits) and visits are thoroughly analyzed to determine their possible relationships and dependency degree. In addition, the be-

[1] http://www.alexa.com/siteinfo/wikipedia.org (Retrieved on 6 June 2012)

[2] http://en.wikipedia.org/wiki/Wikipedia:Academic_studies_of_Wikipedia (Retrieved on 6 June 2012)

havioral conducts expressed through other kinds of requests, such as submit operations or searches, are also taken into account. These kinds of results may be highly valuable in the finding of the type of attention and true impact attracted by Wikipedia, and may even help to explain the origin of certain contributions.

The rest of this paper is structured as follows: first we present some previous studies addressing different topics concerning Wikipedia and, particularly, those related to its utilization by users. Then, the following section describes the data sources used in our analysis and the methodology conducted to perform it. After this, we present our results and conclusions, as well as propose some ideas for further investigation.

2 BACKGROUND

As previously stated, Wikipedia has turned into a prolific research field due to its current popularity and relevance though most of these previous research has focused on quality and evolutionary topics. Wikipedia's underlying approach, that does not rely on any well-known authority for guaranteeing the rigor and veracity of the published information, has made its quality and a promising research area ((Korfiatis et al., 2006), (Giles, 2005), (Chesney, 2006), (Nielsen, 2007)). Other topics in previous investigations regarding Wikipedia have included the reputation of authors (Adler and de Alfaro, 2007) and the differences in evolution tendencies of the different Wikipedia editions ((Capocci et al., 2006) and (Zlatić et al., 2006)). In this way, the number and growth tendency of Wikipedia's articles, authors and visits have been analyzed in several studies (Voss, 2005), (Ortega et al., 2007), (Tony and Riedl, 2009).

The study of the use of Wikipedia has been addressed in the past under very different perspectives. For example, surveys have been used as the main data source in previous analysis such as (Konieczny,), (Schweitzer, 2008), (Waters, 2007) and (Willinsky, 2007). However, these surveys were performed on considerably reduced populations, usually belonging to academic environments and, thus, not representative of general users. In addition, covered topics were not importantly high and limited to the ones implemented in questions.

Another approach significantly different from surveys is the one based on the analysis of users' requests, normally through some of kind of registered log information. This is the basis of studies such as (Urdaneta et al., 2007), (Reinoso et al., 2008), (Reinoso et al., 2009), (Reinoso et al., 2010) or (Reinoso, 2011) that address much more different ways of interaction between Wikipedia and its users. In this same line, our data source consists in a sample of users' requests that have been are registered by Wikimedia Foundation's special Squid servers once they have been conveniently answered. The main features distinguishing our analysis consist in the choice of the most significant Wikipedia editions, regarding both their traffic volumes and their number of articles, and in the large time period considered which covers the whole year 2009.

3 METHODOLOGY

The analysis described in this paper is based on a 1/100 sample taken from the log lines registered by the Wikimedia Foundation's Squid servers every time a user request has been properly served. Log lines included in our sample do not only correspond to Wikipedia, but also to the other wiki-based projects currently maintained by the Wikimedia Foundation. In addition, the sample we have used for this work corresponds to the whole year 2009 and, in total, contains approximately 14,000 million lines. The sample is made up of log lines received in a central aggregator system used to collect and process all of them. This guarantees that our lines correspond to requests made by users all over the world and that they are not affected by the particularities of specific editions.

Log lines used in this study have been registered by special Squid servers. These systems work as reverse proxy servers performing web caching, and have been arranged by the Wikimedia Foundation in order to deal with all the incoming traffic directed to its several projects. Basically, their main purpose is to answer users' requests using their cached contents and to avoid the operation of any other server system placed behind them. This reduces considerably their overload and results in an increase of the overall performance. It is important to note that not all the contents are cacheable, only those requested by non-logged users that share the same HTML code. Log lines from the Wikimedia Foundation Squid systems are finally packed and sent to our systems through a UDP streaming.

Once the log lines have been conveniently stored in our facilities, they can can be analyzed using the JAVA tool we developed for this aim: The *WikiSquilter project*[3]. The analysis consists in a three-step characterization process: parsing, filtering and storage. First, log lines are parsed to extract the fields that provide useful information about users' requests. Then,

[3] http://sourceforge.net/projects/squilter (Retrieved on 14 June 2012)

these information elements are filtered to verify if the corresponding requests comply with the established criteria to be considered of interest for the analysis. Finally, information fields from requests that meet the defined criteria are normalized and stored in a relational database.

As previously mentioned, the log lines we receive correspond to all the projects supported by the Wikimedia Foundation. As we are only interested in those directed to Wikipedia, log lines targeting other projects will be discarded. Furthermore, we pursued that our analysis involved mature and stable editions of Wikipedia. For this reason, we have considered only the requests to the top-ten largest editions, considering both articles and visits. These editions are the German, English, Spanish, French, Italian, Japanese, Dutch, Polish, Portuguese and Russian ones.

Log lines allow to obtain significant information about users' requests, such as the date in which they were sent, or if they caused a write operation into the database. However, most of the data involved in the characterization of requests had to be extracted from the corresponding URLs through a parsing process. This process aimed to determine:

1. The targeted Wikimedia Foundation project (Wikipedia, Wikiversity, Wiktionary, ...)
2. The language edition of the project.
3. If the URL requests an article, its namespace and title.
4. The requested action (edit, submit, history review...) (if any).
5. If the URL corresponds to a search request, the searched topic

Because we aim to study the interaction between users and Wikipedia, we will focus on certain actions requested by them. Particularly, we will look for article visits, contributions (edits), edit requests, submits for previewing and comparisons purposes, historical queries and search operations. Visits to articles are requests dedicated simply to obtain the pages with their contents. Edit operations or contributions are those intended to modify the information presented in the articles and result in issuing write operations to the database servers. In turns, edit requests are sent when users follow the "edit" tab placed on the top-right side of the articles' pages. As a result, users receive the *wikitext* in which the article is stored inside a basic editor that allows to perform the desired changes. Submit operations are those directed to preview the results of the modifications carried out on the current content of an article or to highlight the differences introduced by a given edit operation in course. History queries present the different revisions (edit operations) performed on the contents of an article and which have led to its actual version. Finally, search operations consist of requests for articles containing in their titles a given word or a set of them.

Regarding the implementation aspects, the parser relies on the use of regular expressions to determine the syntactical structure of the URLs. After this, the information components are obtained using string functions. On the other hand, the application's filter checks whether these information elements have been indicated as being of interest to the analysis. To do so, it uses a special hash structure that entails all the specific elements, languages, namespaces, actions, and so forth, that are considered meaningful for the analysis. Apart from these particular elements themselves, the filter also stores their corresponding normalized database code. This way, if a certain element is found in the structure, meaning that it is considered of interest, its database code for the subsequent insert operation to the database can be automatically obtained. The filter has to be queried for each of the information fields parsed from all the processed URLs, so it has to be absolutely accurate and efficient. To achieve an adequate performance concerning this subject, special efforts have been dedicated to reduce the filter's complexity to a O(1) constant level.

The normalized information from users' requests, once stored in the database, will be ready to be used in statistical examinations that aim to determine the degree of relationship between several pairs of measurements. To accomplish this goal, we will apply a test consisting in the calculation of the Pearson's product moment correlation coefficient for the two compared sets of values. This coefficient takes values in the range $[-1,1]$ where closeness to 1 means highly related measurements and 0 indicates no association. The Pearson's product moment correlation coefficient (r) can be computed using the following expression:

$$r = cor(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

The dependence degree between the considered measurements will be analyzed using the correlation of the corresponding sets of values throughout the seven days of the week. Therefore, we have grouped the measurements under study among the weekdays of all the whole weeks corresponding to 2009.

4 RESULTS

The results that we are presenting here are basically aimed to analyze the kind of interactions found between Wikipedia and its users. In addition, several patterns related to different kind of observable attitudes are also introduced.

To begin with, the relationship between visits and contributions can be a good indicator of the degree of participation of a given community of users. In this way, Figures 1 and 2 show the positive correlation between visits and edits throughout the days of the week in the German, English, Spanish, Italian and Russian Wikipedias. The rest of Wikipedias present low correlation values between the two types of requests or even negative ones that indicates that the two kind of requests are inversely correlated. This is the case of the Japanese and Dutch Wikipedias where visits and edits follow completely opposed tendencies.

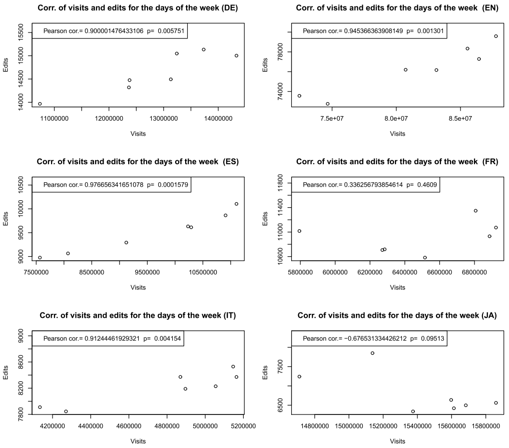

Figure 1: Correlation between visits and edits through the days of the week for the German, English, Spanish, French, Italian and Japanese Wikipedias.

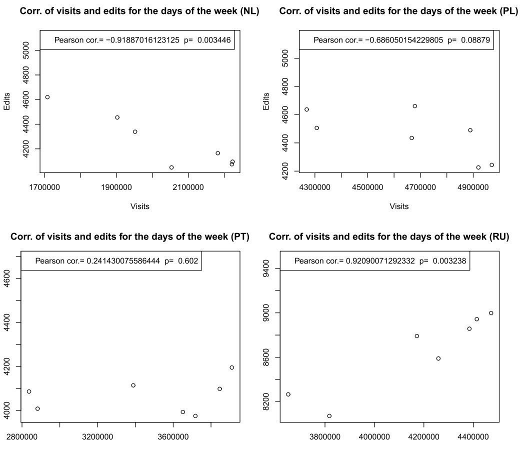

Figure 2: Correlation between visits and edits through the days of the week for the Dutch, Polish, Portuguese and Russian Wikipedias.

If we compare other types of requests to find out if they evolve in a similar way to how visits do, we find that search requests and visits are highly correlated in absolutely all the considered editions. The issue of requests for editing is also positively correlated to visits in all the considered editions. However, history requests are correlated in all the editions except the Japanese one, whereas submit requests are correlated in all the editions except the German and the Japanese ones.

If we focus now on the relationship between edits and requests for editing (Figure 3) we can appreciate that both variables are positively correlated in the German, English, Spanish, Italian and Russian Wikipedias. Interestingly, they are the same Wikipedias in which visits and edits were also correlated. So, we can assume that these editions exhibit massive participation and collaboration of their users on the basis that edits come from the bulk of visits, which means that visitors, at a given moment, turn into contributors. On the contrary, a low correlation between visits and edits may be the result of reluctant-to-contribute attitudes where users massively consult the information offered from the articles, but only a minority of them are responsible for most of the contributions. In other words, editions with low correlations between visits and edits could be supported by an elite of authors.

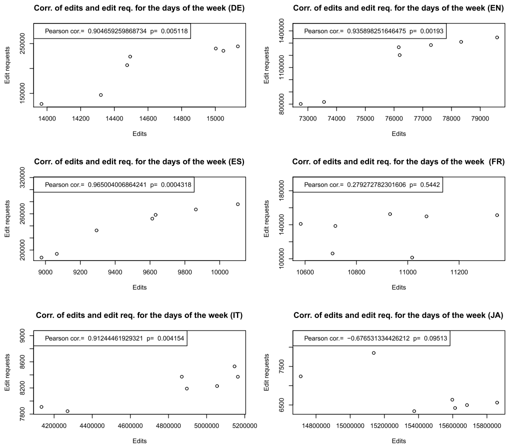

Figure 3: Correlation between edits and edit requests through the days of the week for the German, English, Spanish, French, Italian and Russian Wikipedias.

Regarding the figures about edits and submit requests, we find that only the English, Italian and Russian Wikipedias present correlations between the two measures (Figure 4). That would mean that only the users of these Wikipedias would issue similar values of edits and submit requests in the same days.

In order to deeply address the question of the relationship between visits and edits, we have analyzed the ratio between them for all the considered

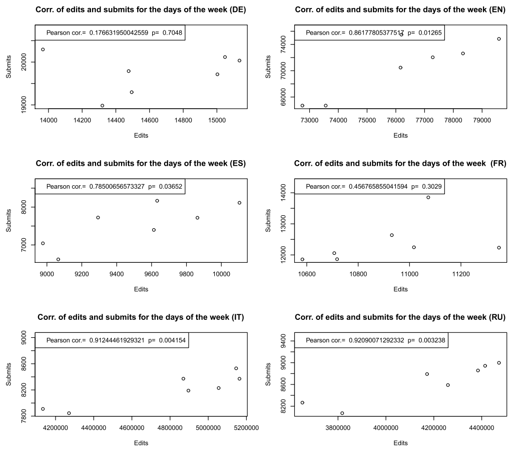

Figure 4: Correlation between edits and submit requests through the days of the week for the German, English, Spanish, French, Italian and Russian Wikipedias.

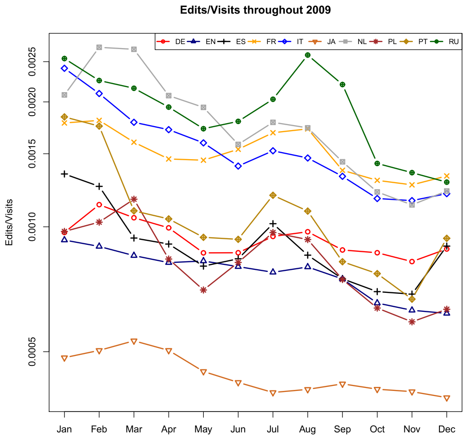

Figure 5: Evolution of the ratio edits to visits throughout 2009 for all the considered Wikipedias.

Wikipedias. Our purpose, in this case, is to assess whether this ratio remains unchanged throughout the year in the different editions and, of course, to determine the editions presenting the highest ratios, as they could be considered as the ones having the most participative communities of users. Thus, Figure 5 presents the evolution of the ratio of edits to visits throughout the entire year. In this figure we can see three groups of editions. The first one is made up of the Dutch, Polish, Italian, French and Russian Wikipedias that present the highest ratios, a second group would consist of the Spanish, Portuguese, English and German Wikipedias with intermediate ratios. Finally, the Japanese is the only one in the third group with the lowest ratio. Interestingly, the Russian and Italian editions, which presented positive correlations between edits and visits, are included among the editions with higher edits/visits ratios. This fact is particularly motivating because it shows how Wikipedias that, presumably, may have an elite of authors present ratios of edits to visits as high as editions purportedly sustained by the whole community of users. Regarding the evolution of the ratio of edits to visits for the different Wikipedia editions, although there are differences in the plots of each one of them, we found a similarities in their shapes. Indeed, most of them decrease, although with different inclines, from January till May-June and they start raising after these two months. Again, there is a general drop after August-September with an slightly increase in December for most of the editions except the Russian, English and Japanese ones. Peaks in the ratios corresponding to summer months may be due to vacancy periods when users have more time to spend contributing that usually. However, a through examination is needed.

Another interesting parameter can be the ratio of edits performed to edits requested, as we have noticed that there is a great number of edit requests that are not finished by the corresponding save operation to the database. This way, Table 1 presents the ratios corresponding to the different editions of Wikipedia in decreasing order. In this case, we did not consider of interest to analyze the evolution of the ratios over time, so we present them aggregated for the entire year. If we compare this table with Figure 5, corresponding to the ratios of edits to visits, we can observe that the Wikipedias having the highest ratios of edits to visits are the ones with the lowest percentages of abandoned edit operations, which is an absolutely interesting finding. The explanation may reside in the fact that there is a kind of editing experience in those editions with higher ratios of edits to visits that result in more completed edit requests.

5 CONCLUSIONS AND FURTHER WORK

After the analysis performed as a part of this work, we can conclude that users from different Wikipedia editions present considerably mismatching behaviors when browsing their contents. One of the more appreciable differences is related to the relationship between visits and contributions (edits). According to our results, the two types of requests are highly correlated throughout the days of the week only for a group of Wikipedias: German, English, Spanish, French, Italian and Russian. This fact can be associated to a more participative attitude on behalf of the users of these editions as it seems that contributions come from the whole mass of visitors. On the contrary, edi-

Table 1: Edit requests finishing with a write operation to the database.

Edition	Edits	Edit requests	Percentage of finished edits
IT	57447	632295	9.09%
FR	76377	941017	8.12%
NL	29799	379450	7.85%
PL	31199	419411	7.44%
RU	60516	814103	7.43%
DE	102442	1426027	7.18%
EN	533879	8026886	6.65%
PT	28469	584498	4.87%
ES	66547	1666890	3.99%
JA	47546	2079305	2.29%

tions where visits and edits are not correlated, or even negatively correlated, can be considered as supported by a minority of contributors. Such a finding may be reinforced by the fact that correlation between edits and requests for editing is not positive but is found on the same Wikipedias. The explanation may reside in the fact that in the editions with an elite of authors, only edits requests coming from its members would be finished with the corresponding contributions saved into the database. As a result, edits cannot be correlated with the general issue of requests for editing.

To go further on the topic, we obtained the ratios of edits to visits for the considered Wikipedias. In fact, we found that communities that supposedly have an elite of authors presented higher ratios. However, two of the editions with high correlation between visits and edits, the Italian and Russian Wikipedias, also presented significantly high values for the considered ratio. After this, we addressed the question of users' reluctance when contributing to their corresponding editions. In this case, we found that the same editions with the highest values of the edits/visits ratios were also the ones having the least number of abandoned edit operations. Therefore, we can conclude that greater number of edits means a kind of expertise and a degree of commitment that result in more finished edits.

We plan to expand this work by also considering the namespaces and topics involved in the different types of requests. In addition, several results of this work, as the double correlation between visits-edits and edits-requests for editing, deserve, in our opinions, thorough efforts. We will also continue to search for a way of relating requests with users, preserving always their fundamental rights for privacy and confidentiality, because any kind of association in this line of research may lead to establish interesting patterns between visitors and contributions.

REFERENCES

Adler, T. B. and de Alfaro, L. (2007). A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, New York, NY, USA. ACM Press.

Capocci, A., Servedio, V. D. P., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., and Caldarelli, G. (2006). Preferential attachment in the growth of social networks: the case of wikipedia.

Chesney, T. (2006). An empirical examination of wikipedia's credibility. *First Monday*, 11(11).

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.

Konieczny, P. Wikis and wikipedia as a teaching tool. *International Journal of Instructional Technology & Distance Learning*, (1).

Korfiatis, Nikolaos, Poulos, Marios, Bokos, and George (2006). Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Information Review*, 30(3):252–262.

Nielsen, F. A. (2007). Scientific citations in wikipedia.

Ortega, F., Gonzalez-Barahona, J. M., and Robles, G. (2007). The top ten wikipedias: A quantitative analysis using wikixray. In *Proceedings of the 2nd International Conference on Software and Data Technologies (ICSOFT 2007)*. INSTICC, Springer-Verlag.

Reinoso, A. J. (2011). *Temporal and behavioral patterns in the use of Wikipedia*. PhD thesis, Universidad Rey Juan Carlos. http://gsyc.es/ ajreinoso/phdthesis.

Reinoso, A. J., Gonzalez Barahona, J. M., Ortega, F., and Robles, G. (2008). Quantitative analysis and characterization of Wikipedia requests. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym '08, New York, NY, USA. ACM.

Reinoso, A. J., González Barahona, J. M., Robles, G., and Ortega, F. (2009). A quantitative approach to the use of the wikipedia. In *ISCC*, pages 56–61. IEEE.

Reinoso, A. J., Ortega, F., Gonzalez-Barahona, J. M., and Herraiz, I. (2010). A statistical approach to the impact of featured articles in wikipedia. In *International Conference on Knowledge Engineering and Ontology Development*, Valencia, Spain.

Schweitzer, N. J. (2008). Wikipedia and psychology: Coverage of concepts and its use by undergraduate students. *Teaching of Psychology*, 35(2):81–85.

Tony, S. and Riedl, J. (2009). Is wikipedia growing a longer tail? In *GROUP '09: Proceedings of the ACM 2009 international conference on Supporting group work*, pages 105–114, New York, NY, USA. ACM.

Urdaneta, G., Pierre, G., and van Steen, M. (2007). A decentralized wiki engine for collaborative wikipedia hosting. In *Proceedings of the 3rd International Conference on Web Information Systems and Technologies*, pages 156–163.

Voss, J. (2005). Measuring wikipedia. In *International Conference of the International Society for Scientometrics and Informetrics : 10th*. ISSI.

Waters, N. L. (2007). Why you can't cite wikipedia in my class. *Commun. ACM*, 50(9):15–17.

Willinsky, J. (2007). What open access research can do for wikipedia. *First Monday*, 12(3).

Zlatić, V., Božičević, M., Štefančić, H., and Domazet, M. (2006). Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(1).