Chapter 8
Behavioral Aspects in the Interaction Between Wikipedia and its Users

Antonio J. Reinoso and Juan Ortega-Valiente

Abstract Wikipedia continues to be the most well-known on-line encyclopedia and receives the visits of millions of users on a daily basis. Its contents correspond to almost all the knowledge areas and are altruistically contributed by individuals and organizations. In addition, users are encouraged to add their own contributions according to the Wikipedia's own supporting paradigm. Its progression to a mass phenomenon has propitiated many studies and research initiatives. Therefore, topics such as the quality of the published contents or the authoring of its contributions have been widely developed. However, very few attention has been paid to the behavioral aspects characterizing the interaction between Wikipedia and its users. Henceforth, this chapter aims to determine the habits exhibited by users when browsing the Wikipedia pages. Particularly, we will focus on visits and contributions, as they constitute the two most common forms of interaction. Our study is based on a sample of the requests submitted to Wikipedia, and its results are twofold: on the one hand, it provides different metrics concerning users' behavior and, on the other, presents particular comparisons among different Wikipedia editions.

Keywords Wikipedia · Use patterns · Behavioral patterns · Traffic characterization

1 Introduction

Wikipedia can be considered 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

A. J. Reinoso (✉) · J. Ortega-Valiente
Department of ICT Engineering, Universidad Alfonso X el Sabio, Vva. de la Cañada, 28691 Madrid, Spain
e-mail: areinpei@myuax.com

J. Ortega-Valiente
e-mail: jvalient@myuax.com

C. Lai et al. (eds.), *Distributed Systems and Applications of Information Filtering and Retrieval*, Studies in Computational Intelligence 515,
DOI: 10.1007/978-3-642-40621-8_8,

construction of any kind of knowledge. The resulting compendium of contents will remain available to 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 puts its website within the six most visited ones in all the Internet.[1]

As a result of such popularity, Wikipedia has turned into a subject of interest for many 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 evolution tendencies. Our work, on the other hand, aims to address the use given to Wikipedia by some of its most notorious communities of users through the analysis of the most common forms of interactions carried out by users.

Thus, in this study we will address several issues related to the use given to the different editions of Wikipedia by their corresponding communities of users. In particular, we will examine users' behavioral habits extracted from the requests they submit when browsing Wikipedia. These habits include both general attitudes, like participation or collaboration, as well as more particular ones, such as the previewing of changes when editing articles or users' reluctance to commit changes at the moment of contributing. Considering that different Wikipedia editions may provide very different user behavioral patterns when examining the forms of interaction with their respective communities of users, we will compare the results obtained for each different edition analyzed and evaluate the differences and similarities found among them.

Our results aim to present observed patterns related to the most common interactions between Wikipedia and some of its most prolific communities of users. In particular, the resulting relationships between contributions (edits) and visits are thoroughly analyzed to present their respective dependency degrees. In addition, the behavioral habits derived from certain measures such as participation, reluctance and, even more, the relationships among them are equally introduced. Finally, 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 finding 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 chapter 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 research.

[1] http://www.alexa.com/siteinfo/wikipedia.org (Retrieved on 6 February 2013)

[2] http://en.wikipedia.org/wiki/Wikipedia:Academic_studies_of_Wikipedia (Retrieved on 6 February 2013)

2 Background

As previously stated, Wikipedia has turned into a prolific research field due to its overwhelming popularity and relevance. Wikipedia's underlying approach, based on free access and contributions from all users on the Internet, does not rely on any well-known authority to check the veracity of the published information, nor does it have any censoring authority, and has therefore made the topic of its quality and reliability a promising research area, where studies such as [1–4] have focused on different ways to evaluate it. Other topics in previous research studies regarding Wikipedia have included the reputation of the authors [5] and the differences in evolution tendencies of its editions [6, 7]. In this way, the number and growth tendency of Wikipedia's articles, authors and types of visits have been analyzed in many studies, being some of the most relevant [8–10].

The study of the use given to Wikipedia has been addressed in the past under many different perspectives. For example, the use of surveys has been the main data source for several previous studies, including [11–14]. However, these surveys were performed on considerably reduced, and very specific, populations, usually belonging to academic environments and, thus, not representative of general users. In addition, the topics covered were not highly important and were limited to the ones specified in the questions included in the surveys.

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 several studies including [15–17], which address much more specific ways of interaction between Wikipedia and its users. In this same line, our data source consists in a sample of the users' requests that have been registered by the Wikimedia Foundation's special Squid servers once they have been conveniently answered. The main features distinguishing our analysis from the rest consist on 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 chapter is based on a sample from the log lines registered by Wikimedia Foundation's special Squid servers every time they properly answer a user request. 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, it contains approximately 14,000 million lines. It is important to note that the log lines comprised in our sample are extracted from a central aggregator system that receives and process the lines generated by all the Squid servers deployed by the Wikimedia Foundation. 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.

The Squid systems that register the log information that we are using for this study work as reverse proxy servers, performing web caching of Wikipedia and other wiki-based initiatives and projects developed by the Wikimedia Foundation. They have been arranged in order to deal with all the incoming traffic directed to them.

Basically, their main purpose consists in answering users' requests using their cached contents to avoid the operation of any other server system placed behind them, specially web servers and database servers. This reduces their overload considerably and results in an increase of the overall performance, as these Squid servers are taking much of the load of the requests directly. It is important to consider that not all Wikipedia contents are cacheable; while standard anonymous users all receive the same HTML content code, registered users' requested pages may contain additional dynamic content (such as personalization options) or metadata, and therefore cannot be cached in intermediate proxy servers. After being sampled by a dedicated service, Wikimedia Foundation Squid log lines are packed and piped to our systems through an UDP streaming.

After receiving these log lines, they are properly stored in our facilities, where they are analyzed using a JAVA-based tool developed for this specific purpose: *The WikiSquilter Project*.[3] The analysis of these log lines 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, 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 requests specifically directed to Wikipedia, log lines targeting other projects are, therefore, discarded. Furthermore, our analysis involves only mature and stable editions of Wikipedia; reason why we have considered requests made only to the top-ten largest editions, considering both articles and visits. The top ten editions which meet these criteria are the German, English, Spanish, French, Italian, Japanese, Dutch, Polish, Portuguese and Russian ones.

Log lines allow us to obtain significant information about users' requests, including 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 those requests had to be extracted from their corresponding URLs through an advanced parsing process. This process aims to determine and classify these requests, to be able to ignore those which are not relevant for this study:

1. The targeted Wikimedia Foundation project (Wikipedia, Wikiversity, Wiktionary, ...).

[3] http://sourceforge.net/projects/squilter (Retrieved on 14 February 2013)

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), requests for editing, submits for previewing changes and comparisons purposes, historical queries and search operations. Visits to articles are requests dedicated simply to obtain the pages with their contents to visualize them. 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 turn, requests for editing 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 them 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 current 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 level 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 sets of measures. 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 proximity to 1 means highly related measurements and to 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 dependency degree between some of the considered measures will be analyzed using the correlation of the corresponding sets of values throughout the 7 days of the week. Therefore, we have grouped the measurements under study among the weekdays for all the weeks corresponding to 2009.

4 Results

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

To begin with, the relationship between visits and contributions can be considered as a good indicator of the degree of participation of a given community of users. In this way, Fig. 1 shows the correlation obtained between visits and edits throughout all the days of the week in the German, English, Spanish, French, Italian and Japanese editions of Wikipedia, while in Fig. 2 the same correlation between visits and edits is presented but for the Dutch, Polish, Portuguese and Russian editions of Wikipedia. The results clearly show a highly positive correlation (over 0.9) between edits and visits in the German, English, Spanish, Italian and Russian editions. In contrast, the Dutch edition presents a high negative correlation and the Japanese and Polish editions a medium negative correlation; this indicates that in these three editions, an inverse correlation was found, as visits and edits follow completely opposed tendencies. In the case of the French and Portuguese editions, high p-values do not allow to pronounce about requests being correlated.

When we compared other types of requests to find out whether they evolved in a similar way as visits do, we found that search requests and visits are highly correlated in all ten considered editions (German, English, Spanish, French, Italian, Japanese, Dutch, Polish, Portuguese and Russian) showing correlation coefficients over 0.9. Figure 3 presents the correlation graphs for the six first editions aforementioned. In the same way, requests for editing are correlated to visits for all the considered editions.

Moreover, when calculating the correlation between history requests and visits, we observed that the requests were positively correlated for all the considered editions except the Japanese one. Figure 4 shows the graphs corresponding to five of the positively correlated editions and to the Japanese one. When analyzing submit requests an visits, we found that the English, Spanish, Italian, Dutch, Polish, Portuguese and Russian presented positive correlations. The French edition, in turn, only showed a medium positive correlation value (barely over 0.5), and both the German and Japanese editions displayed no correlation at all. Figure 5 shows three of the editions in which visits and submit requests were positively correlated (Eng-

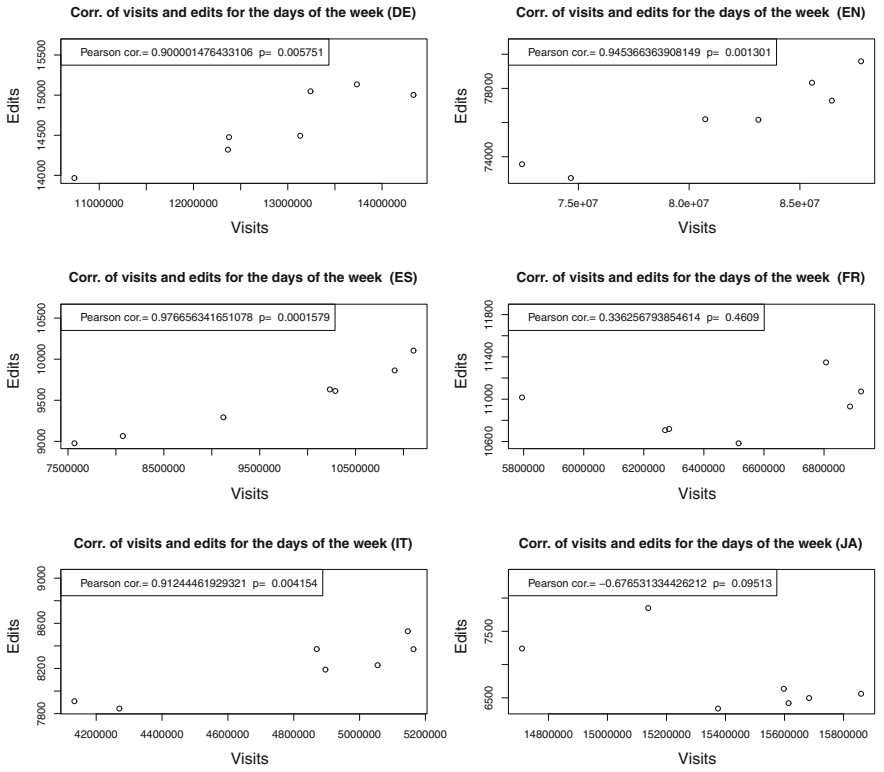

Fig. 1 Correlation between visits and edits through the days of the week for the German, English, Spanish, French, Italian and Japanese Wikipedias

lish, Spanish and Italian) as well as the correlations obtained for the French, German and Japanese editions.

If we focus now on the relationship between edits and requests for editing (Fig. 6) we can appreciate that both variables are positively correlated in the German, English, Spanish and Italian editions. In the case of the Japanese edition, a negative correlation was found. The high French edition's p-value does not allow to pronounce about the correlation of its requests. Interestingly, Wikipedias where edits and requests for editing were correlated are the same on 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 are most likely supported by a reduced elite of authors.

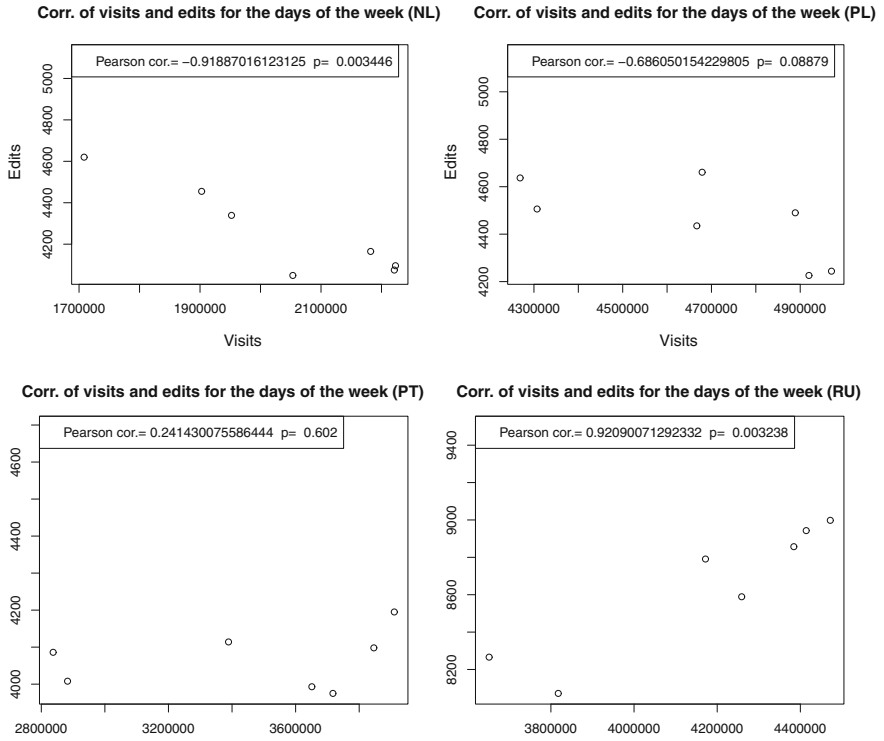

Fig. 2 Correlation between visits and edits through the days of the week for the Dutch, Polish, Portuguese and Russian Wikipedias

Regarding the correlation between edits and submit requests, we found that only the English, Spanish, Italian and Russian Wikipedias present positive correlations between the two measures (Fig. 7). That would mean that only the users of these Wikipedias would issue similar values of edits and submit requests in the same days, which may be related to attitudes in favor of checking the introduced changes as a previous step to submit them. Both French and German editions' respective values prevent any pronunciation about this type of requests.

In order to properly address the question of the relationship between visits and edits, we have analyzed the ratio between them for all the considered 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 which editions present the highest ratios, as they could be considered as the ones having the most participative communities of users. Thus, Fig. 8 presents the evolution of the ratio of edits to visits throughout the entire year for the ten Wikipedia editions selected. In this figure we can see three groups of editions. The first one is formed up by the Dutch, Polish, Italian, French and Russian editions that present the highest ratios; the second group which consists of the Spanish, Portuguese, English and German editions with

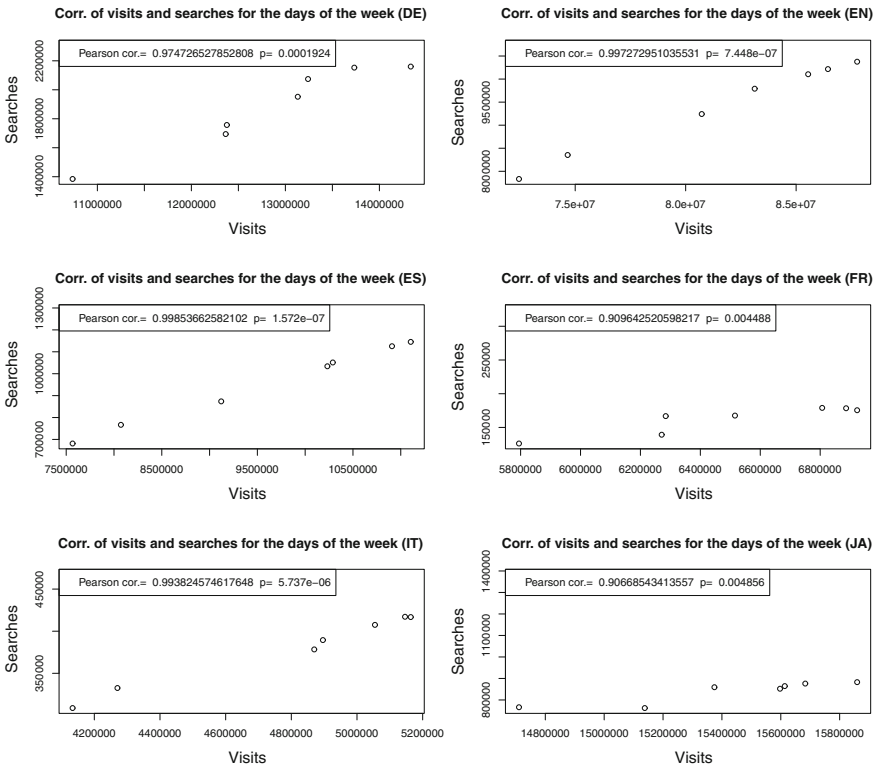

Fig. 3 Correlation between visits and search requests through the days of the week for the German, English, Spanish, French, Italian and Japanese Wikipedias

intermediate ratios; and finally, the Japanese edition alone forming 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 to visits ratios. This fact is particularly interesting because it shows how Wikipedias that, in theory, would be sustained by the whole community of users present ratios of edits to visits as high as editions potentially supported by an elite of authors. 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 follow a decreasing start from January till May–June, an increase trend lasting the two following months to then return to the initial decreasing trend up to December, when some of the editions experienced an small increase trend again, with the exception of the English, Japanese and Russian ones. Most of the increase peaks found correspond to summer months, and may very well be connected to the fact that users tend to have more free time in this period and therefore may have more time to contribute. However, more data would be required to confirm whether this connection is accurate or not.

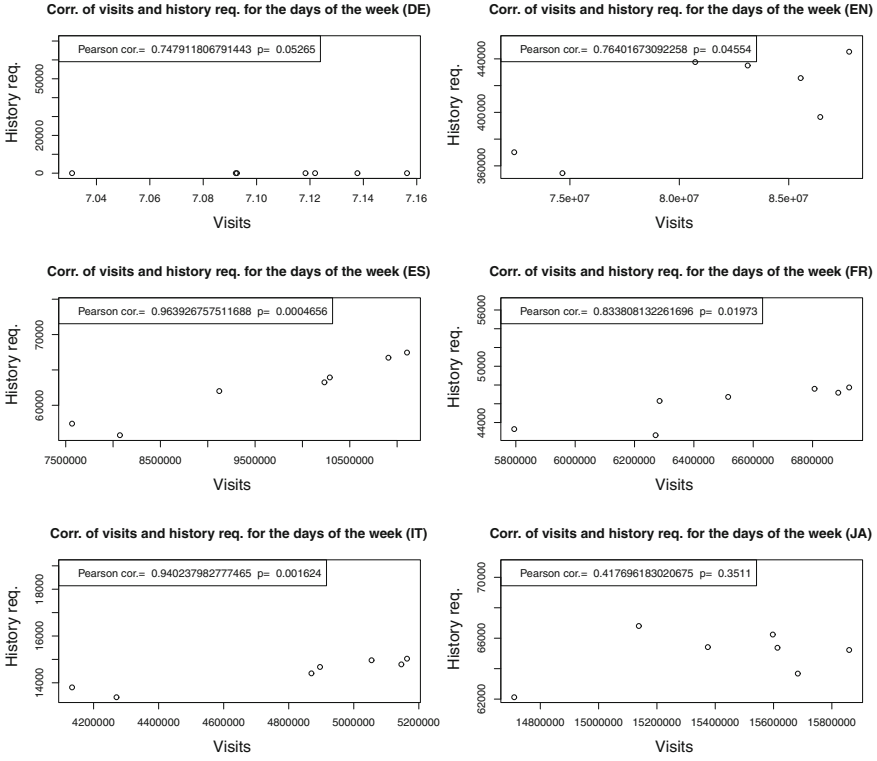

Fig. 4 Correlation between visits and history requests through the days of the week for the German, English, Spanish, French, Italian and Japanese Wikipedias

Another interesting parameter evaluated as a part of this study is 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 their corresponding save operations to the database (that would make an actual contribution). This way, Table 1 presents the percentages of finished contributions corresponding to the different editions decreasingly ordered. In this case, it was not found of relevance to analyze the evolution of the ratios over time, so we presented them aggregated for the entire year. If we compare this table with Fig. 8, which corresponds to the ratios of edits to visits, we can observe that the Wikipedias having the highest ratios of edits to visits match the ones with the lowest percentages of abandoned edit operations, which is, in fact, 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 requests for editing.

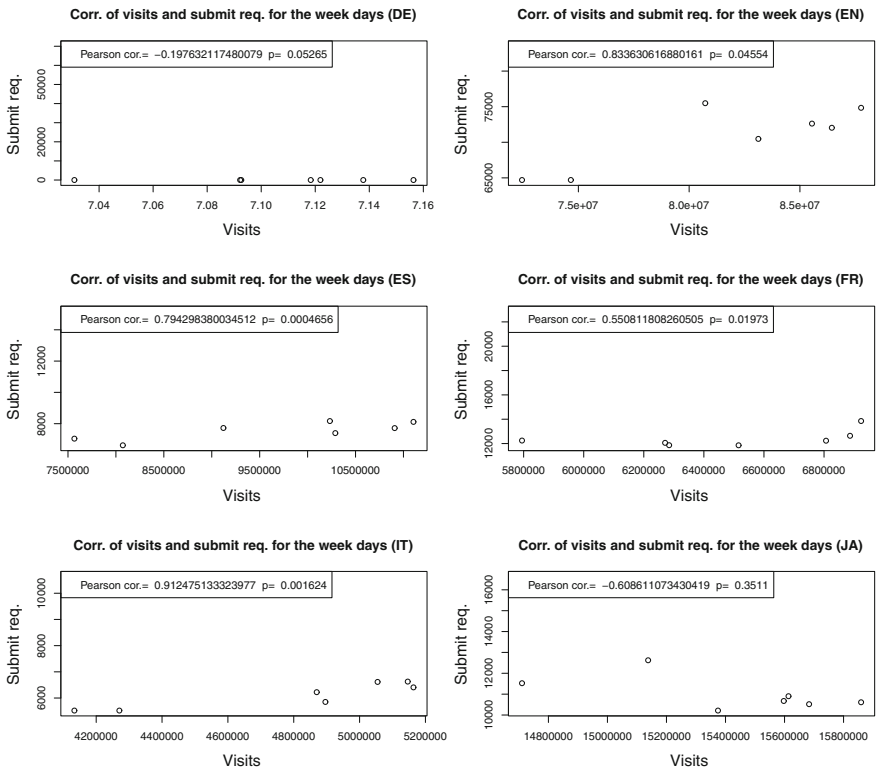

Fig. 5 Correlation between visits and submit requests through the days of the week for the German, English, Spanish, French, Italian and Japanese Wikipedias

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 different 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 the following Wikipedia editions: German, English, Spanish, French, Italian and Russian. This fact can be associated to a more participative attitude of the users of these editions, as it seems that contributions come from the whole mass of visitors. On the contrary, editions 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 again not positive for these editions. The explanation may reside in the fact that in these editions, as an elite of authors would be responsible for the majority of contributions,

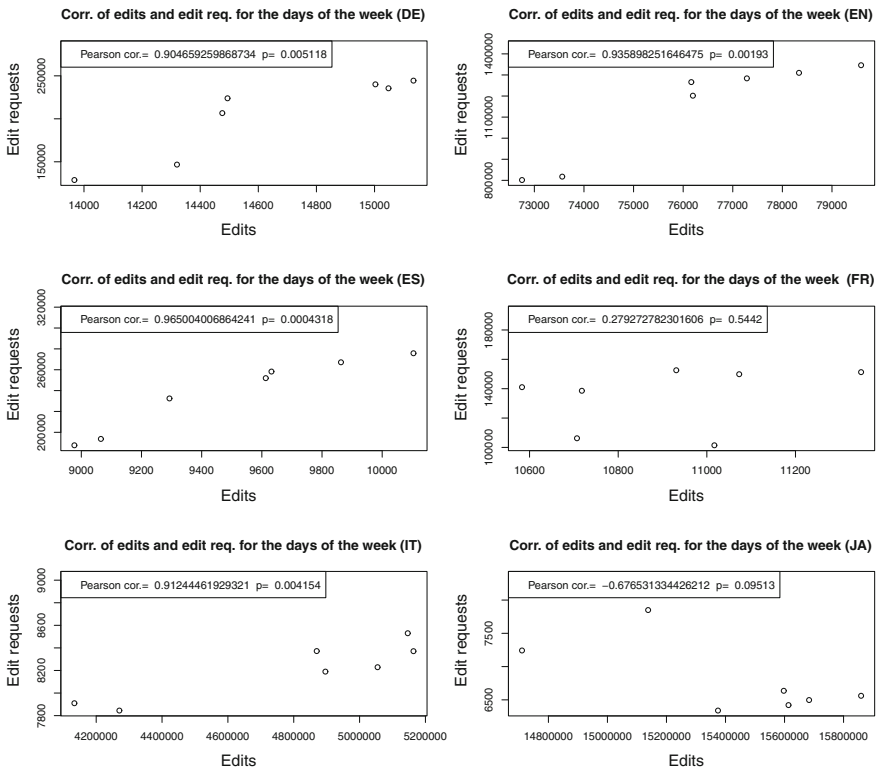

Fig. 6 Correlation between edits and requests for editing through the days of the week for the German, English, Spanish, French, Italian and Japanese Wikipedias

only edits coming from them would be appropriately finished whilst the rest would be abandoned.

To get further insight on the topic, we obtained the ratios of edits to visits for the considered Wikipedia editions. 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.

Among the possible expansions that can arise for this work, we are more inclined to continue by taking into consideration the namespaces and topics involved in the different types of requests evaluated. Furthermore, several results of this work, and specially the correlation found in both visits-edits and edits-requests for editing,

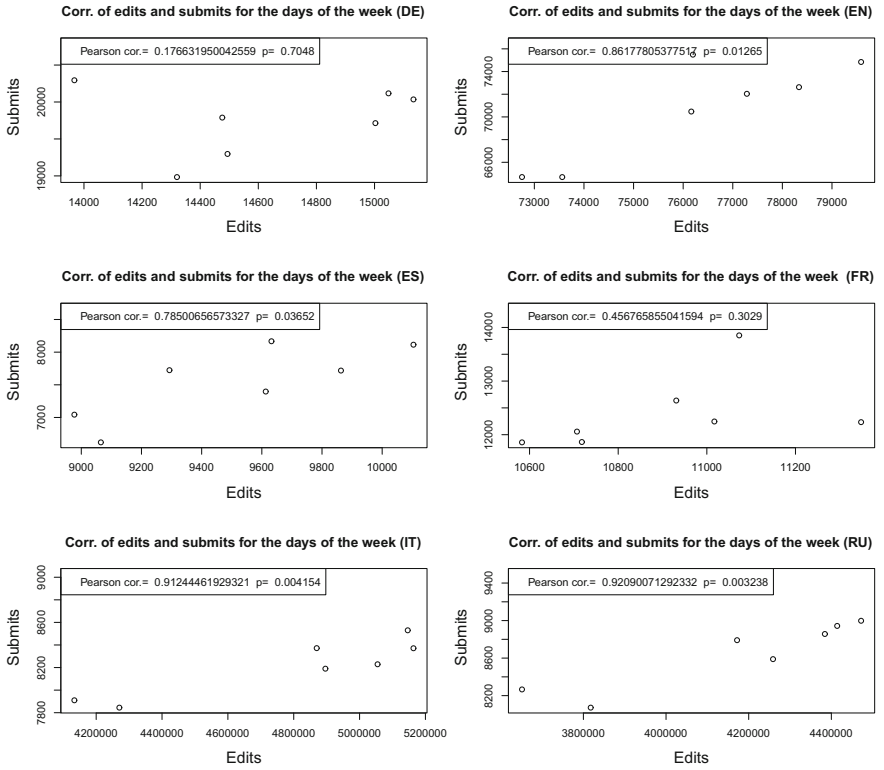

Fig. 7 Correlation between edits and submit requests through the days of the week for the German, English, Spanish, French, Italian and Russian Wikipedias

Table 1 Requests for editing completed (i.e. finished by a write operation to the database)

Edition	Edits	Edit requests	Percent of finished edits (%)
Italian(IT)	57,447	632,295	9.09
French(FR)	76,377	941,017	8.12
Dutch (NL)	29,799	379,450	7.85
Polish (PL)	31,199	419,411	7.44
Russian (RU)	60,516	814,103	7.43
German (DE)	102,442	1,426,027	7.18
English (EN)	533,879	8,026,886	6.65
Portuguese (PT)	28,469	584,498	4.87
Spanish (ES)	66,547	1,666,890	3.99
Japanese (JA)	47,546	2,079,305	2.29

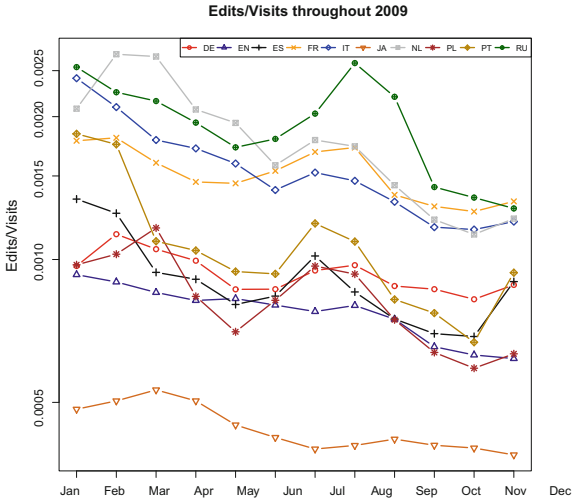

Fig. 8 Evolution of the ratio edits to visits throughout 2009 for all the considered Wikipedias

present a perfect case for further study and for a more thorough comparison. We also intend to 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 could potentially lead to establishing interesting usage patterns between visitors and contributors as well as to enable some form of user profiling.

Moreover, another possible expansion of this work is to analyze a larger sample of the logs to verify the accuracy of the tendencies found in this study both in edit and visit requests, and whether this tendency is stable or varies though different periods of time. This could lead to define if the visits and edits to the Wikipedia articles, in the ten selected editions, grow steadily or not, and find out if the are differences between the tendencies of finished and unfinished edits. Another possible variation would be to increase the number of editions included, duplicating it for example, and checking if they follow similar usage tendencies to the top ten ones.

References

1. Korfiatis, N., Poulos, M., Bokos, G.: Evaluating authoritative sources using social networks: an insight from Wikipedia. Online Inf. Rev. **30**(3), 252–262 (2006)
2. Giles, J.: Internet encyclopaedias go head to head. Nature **438**(7070), 900–901 (2005)
3. Chesney, T.: An empirical examination of Wikipedia's credibility. First Monday **11**(11), (November 2006)
4. Nielsen, F.A.: Scientific citations in Wikipedia. First. Monday **12**(8), (May 2007)
5. Adler, T.B., de Alfaro, L.: A content-driven reputation system for the Wikipedia. In: WWW '07: Proceedings of the 16th International Conference on World Wide Web, pp. 261–270. ACM Press, New York (2007)

6. Capocci, A., Servedio, V.D.P., Colaiori, F., Buriol, L.S., Donato, D., Leonardi, S., Caldarelli, G.: Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. Phys. Rev. E. **74**(3). doi:10.1103/PhysRevE.74.036116. http://link.aps.org/doi/10.1103/PhysRevE.74.036116 (2006)
7. Zlatić, V., Božičević, M., Štefančić, H., Domazet, M.: Wikipedias: collaborative web-based encyclopedias as complex networks.Phys. Rev. E Stat. Nonlin. Soft. Matter Phys. **74**(1), (2006)
8. Voss, J.: Measuring Wikipedia. In: 10th International Conference of the International Society for Scientometrics and Informetrics (ISSI), (2005)
9. Ortega, F., Gonzalez-Barahona, J.M., Robles, G.: 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 (2007)
10. Tony, S., Riedl, J.: Is Wikipedia growing a longer tail? In: GROUP '09: Proceedings of the ACM: International Conference on Supporting Group Work, pp. 105–114. ACM, New York, 2009
11. Konieczny, P.: Wikis and Wikipedia as a teaching tool. Int. J. Instr. Technol. Distance Learn. **1**, 15–34 (2007)
12. Schweitzer, N.J.: Wikipedia and psychology: coverage of concepts and its use by undergraduate students. Teach. Psychol. **35**(2), 81–85 (2008)
13. Waters, N.L.: Why you can't cite Wikipedia in my class. Commun. ACM **50**(9), 15–17 (2007)
14. Willinsky, J.: What open access research can do for Wikipedia. First Monday **12**(3), (March 2007)
15. Urdaneta, G., Pierre, G., van Steen, M.: A decentralized wiki engine for collaborative Wikipedia hosting. In: Proceedings of the 3rd International Conference on Web Information Systems and Technologies, pp. 156–163 (2007)
16. Reinoso, A.J., Ortega, F., Gonzalez-Barahona, J.M., Herraiz, I.: A statistical approach to the impact of featured articles in Wikipedia. International Conference on Knowledge Engineering and Ontology Development, Valencia (2010)
17. Reinoso, A.J.: Temporal and behavioral patterns in the use of Wikipedia. Ph.D. thesis, Universidad Rey Juan Carlos (2011). http://gsyc.es/ ajreinoso/phdthesis