Temporal characterization of the requests to Wikipedia

Antonio J. Reinoso, Jesus M. Gonzalez-Barahona, Rocio Muñoz-Mansilla and Israel Herraiz

Abstract This paper presents an empirical study about the temporal patterns characterizing the requests submitted by users to Wikipedia. The study is based on the analysis of the log lines registered by the Wikimedia Foundation Squid servers after having sent the appropriate content in response to users' requests. The analysis has been conducted regarding the ten most visited editions of Wikipedia and has involved more than 14,000 million log lines corresponding to the traffic of the entire year 2009. The conducted methodology has mainly consisted in the parsing and filtering of users' requests according to the study directives. As a result, relevant information fields have been finally stored in a database for persistence and further characterization. In this way, we, first, assessed, whether the traffic to Wikipedia could serve as a reliable estimator of the overall traffic to all the Wikimedia Foundation projects. Our subsequent analysis of the temporal evolutions corresponding to the different types of requests to Wikipedia revealed interesting differences and similarities among them that can be related to the users' attention to the Encyclopedia. In addition, we have performed separated characterizations of each Wikipedia edition to compare their respective evolutions over time.

Antonio J. Reinoso
LibreSoft Research Group (URJC), C/ Tulipan s/n, 28933, Mostoles, Madrid, Spain, e-mail: ajreinoso@libresoft.es

Jesus M. Gonzalez-Barahona
LibreSoft Research Group (URJC), C/ Tulipan s/n, 28933, Mostoles, Madrid, Spain, e-mail: jgb@libresoft.es

Rocío Muñoz-Mansilla
Department of Automation and Computer Science (UNED), C/ Juan del Rosal, 16, 28040, Madrid, Spain, e-mail: rmunoz@dia.uned.es

Israel Herraiz
Department of Applied Mathematics and Computing (UPM), C/ Profesor Aranguren s/n, 28040, Madrid, Spain, e-mail: israel.herraiz@upm.es

1 Introduction

Wikipedia continues to be an absolute success and stands as the most relevant wiki-based platform. It provides a rich set of contents belonging to every knowledge area that are offered in different formats that range from text to multimedia resources. In addition, the Wikipedia's supporting paradigm, which is based on individuals collaboration and joint of efforts to produce and contribute pieces of knowledge that will remain available for the whole community. The consolidation of Wikipedia as a reference tool and a platform for mass collaboration is endorsed by the increasing number of visits to its portal. In fact, the Wikipedia domain remains within the six most visited ones all over the Internet.

Wikipedia is divided in 268 [1] editions corresponding each to a different language and its overall relevance can be simply measured in terms of the number of visits it receives. Currently, the overall set of Wikipedias editions are receiving approximately 13,500 million visits a month. This constitutes an absolute challenge in terms of management of requests and content delivery. On the other hand, Wikipedia organizes the information it offers in encyclopedic entries commonly referred as articles. At the moment of writing this paper, the different Wikipedia editions add up to almost 18 million articles and this number does not stop growing.

As a result of this relevance, Wikipedia has evolved into a subject of increasing interest for researchers [12]. In this way, quantitative examinations about its articles, authors, visits or contributions have made part of different studies [11, 6, 3]. However, most of previous research involving Wikipedia is concerned with the quality and reliability of its contents ([2, 1] or [7, 5, 4]) or focus on the study of its growth tendency and evolution [9, 8]. By contrast, very few studies [10] have been devoted to analyze the manner in which users interact and make use of Wikipedia.

Therefore, this paper presents an empirical study encompassing a temporal characterization that may help to describe the evolution over time of users' interactions with Wikipedia. Furthermore, we will compare the results obtained for the different editions in order to analyze the main differences and similarities among them.

Our analysis focuses on the most relevant Wikipedia editions in terms of their volumes of articles and number of traffic. In addition, the period of time considered correspond to a whole year (2009). Our main data source consists in users' requests to Wikipedia previously stored by special servers deployed to deal with the incoming traffic. Information about each individual request is registered in the form of a log line whose fields are processed by an ad-hoc developed application. This application filters the requests considered of interest for our analysis and stores its information elements into a database for further examinations.

The rest of the paper is structured as follows: first of all, we describe the data sources used in our analysis as well as the methodology followed to conduct our work. After this, we present our results and, finally, we present our conclusions and propose some ideas for further work.

[1] `http://stats.wikimedia.org/EN/Sitemap.htm`

2 The data sources

This section aims to describe the information sources involved in our study and used as the main data feeding to perform our analysis. The visits to Wikipedia, in a similar way to any other Internet site, are issued in the form of URLs sent from the users' browsers. These URL's are registered by the Wikimedia Foundation Squid servers in the form of log lines after serving the requested content.

Therefore, the following sections present the principal aspects related to how the Squid log lines used in this analysis are registered, their way to our storage systems and the most important information elements that they contain.

2.1 The Wikimedia Foundation Squid subsystem

Squid servers are usually used to perform web caching working as proxy servers. In this way, they can cache the contents browsed by a group of users to make them available for further requests. This results in an important decrease of the bandwidth consumption and in a more efficient use of the network resources. Furthermore, Squid servers may be used to speed up web servers by caching the contents requested repeatedly to them. Under this approach, Squid servers are said to work as reverse proxy servers because they try to reply to the received requests using the cached contents, what reduces, if so, the workload of both the web and database servers placed behind them.

The Squid operation is based on web caching and, hence, it is aimed to avoid the participation of the other database and web server systems in operations for serving requested contents. In this way, when a requested page can be found on a Squid server and it is up-to-date, the page is directly served from the Squid and neither the database server nor the web server have to be involved in the delivery process. Otherwise, the request is sent to the web servers which elaborate the corresponding HTML code and submit it to the Squid for its caching and final delivery to the user.

As the Wikimedia Foundation maintains several wiki-based projects, such us Wikipedia, Wikiversity or Wikiquote, the Squid layers have to deal with all the traffic directed to these projects. As a part of their job, Squid systems do log information about every request they serve whether the corresponding contents stem from their caches or, on the contrary, are provided by the web servers. In the end, Squid systems register a log line with different kind of information for each served request and these lines can be written to a file or sent to another process through a pipe as in the case of the Wikimedia Foundation.

Each log line from a Wikimedia Squid server corresponds to a served user request and constitute a really valuable feed because, among several other information, it includes the URLs submitted by the user along with the date at witch the corresponding content was sent in response.

3 Methodology

The analysis presented here is based on a sample of the Wikimedia Foundation Squid log lines corresponding to the entire year 2009. The sampling factor used has been $1/100$, so this study has included the characterization of the 1% of the overall traffic directed to all the projects maintained by the Wikimedia Foundation during the whole year 2009. In general terms, more than 14,000 million log lines have been parsed and filtered for this study.

This analysis has focused just on the traffic directed to the Wikipedia project and to ensure that the study involved mature and highly active language editions, only the requests corresponding to the ten most visited ones have been considered. These editions are the German, English, Spanish, French, Italian, Japanese, Dutch, Polish, Portuguese and Russian ones.

Once the log lines from the Wikimedia Foundation Squid systems have been received in our facilities and conveniently stored, they become ready to be analysed by the tool developed for this aim: The WikiSquilter project. The analysis consists on a characterization based on a parsing process to extract the relevant elements of information prior to a filtering one according to the study directives. As a result of both processes, necessary data to conduct a characterization are obtained and stored in a relational database for further analysis.

The lines received from the Wikimedia Foundation offer a valuable information source but they do not include specific information elements to describe certain characteristics of the corresponding requests. However, these elements can be obtained from the URL embedded in each line which, therefore, has to be parsed looking for specific data serving as characterization elements.

More in the detail, the application parser is devoted to determine the following information elements:

1. The Wikimedia Foundation project, such us Wikipedia, Wiktionary or Wikiquote, to which the URL is directed.
2. The corresponding language edition of the project.
3. When the url requests an article, its namespace.
4. The action (edit, submit, history review...) requested by the user (if any).
5. If the URL corresponds to a search request, the searched topic
6. The title of every requested article or user page name.

The parsing process relies on the use of regular expressions for verifying whether an URL, or a part of it, matches a given pattern. If so, its components can be obtained using common functions for string manipulation. On the other side, the filter process consists in assessing whether an URL has to be considered significant for our analysis according to its directives. This is accomplished by checking whether the information elements it contains, once parsed, has been indicated to be filtered.

The application has been designed and developed according to the principles of efficiency, robustness and accuracy. However, flexibility and extensibility guidelines have been also strictly followed. Efficiency has been achieved through several elements such as multithreaded design and filter's O(1) complexity derived of the use of hash tables. The application robustness has been implemented by URL malformation detection and preprocessing to avoid non-appropriate characters and has allowed to rightly process all the log lines to be analyzed. Flexibility makes the application ready to be used with any sort of log files just specifying in a XML log files the elements to be parsed and filtered. The software architecture of the application allows to easily include new services that can even involve new data to be processed, so extensibility has been also considered.

4 Analysis and results

In the following we are presenting our most important results about the temporal characterization of users' requests submitted to Wikipedia. First of all, we analyze if the traffic to Wikipedia can reliably model the overall traffic to the Wikimedia Foundation. After this, we compare the evolution of the different types of requests over time. Concerning this topic, we will present the different patterns found, paying special attention to the ones showing repetitive schemes. This examination has been specially conducted under a comparative approach to determine whether or not the same tendencies are maintained in every considered Wikipedia edition. Finally, our analysis allow to obtain valuable information about the ratios corresponding to the different types of requests that is also presented.

4.1 The traffic to Wikipedia as a model of the traffic to the Wikimedia Foundation

Figure 1 presents the yearly evolution of the traffic directed to the aggregated set of the editions of Wikipedia in order to compare it with the overall traffic directed to all the projects maintained by the Wikimedia Foundation. Moreover, Figure 1 also plots the number of requests filtered after our analysis. As we can see, all three lines, each in its corresponding scale, present a relative similar behavior over time. The decrease appreciated since November till the end of the year is documented in [2] and is due to a problem in the reception of the UDP packets. The slumps in the number of visits that appear in February, June, July and October correspond to the days in which we were not able to receive and store the log lines from the Wikimedia Foundation Squid systems due to technical problems related to our system's storage capacity.

In order to examine more accurately the relationship between the traffic to Wikipedia and to all the Wikimedia Foundation projects, Figure 2 shows the correlation between the daily measures of both traffics corresponding to the entire year. As it is shown, there is a positive correlation between the two variables so, effectively,

[2] `http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm`

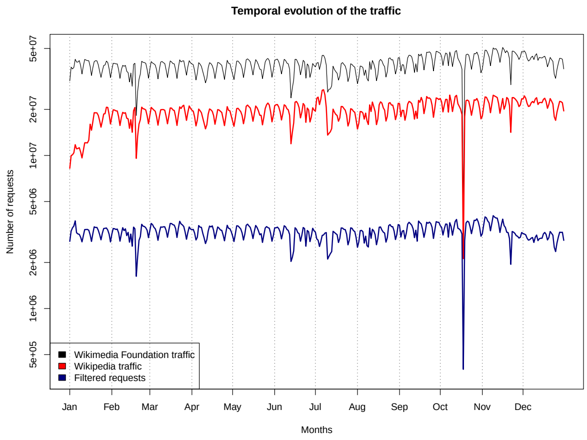

Fig. 1 Evolution of the traffic throughout 2009.

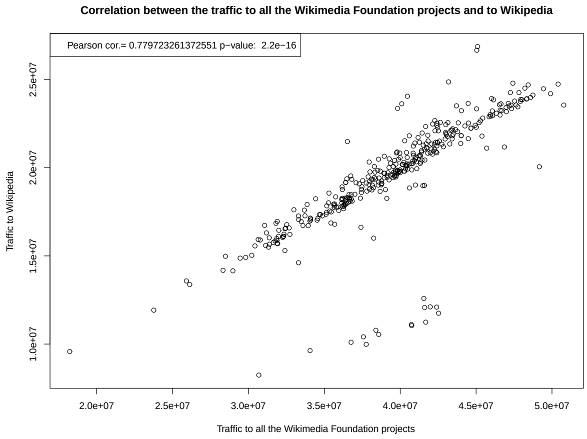

Fig. 2 Correlation between the traffic to Wikipedia and to the whole set of Wikimedia Foundation projects throughout 2009.

Wikipedia traffic can serve as model of the overall traffic to different Wikimedia Foundation projects.

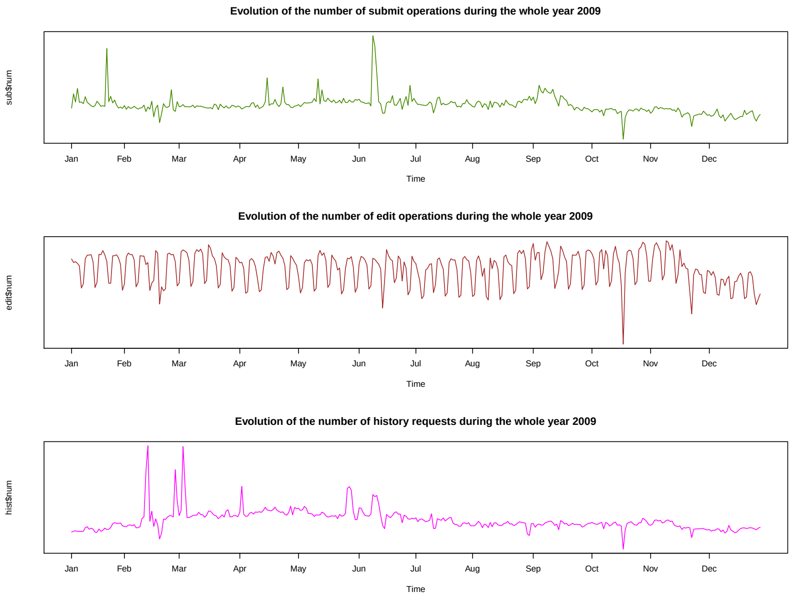

Fig. 3 Evolution of submits, edit requests and history reviews throughout 2009.

4.2 *Temporal evolution of the different types of requests to Wikipedia*

If we separate the requests to Wikipedia according to their types, Figures 3 and 4 show how each one of them evolves throughout the entire year 2009. We are considering a visit to an article as its page request for reading and without involving any other action. In turn, edit operations are intended as modifications over the content of articles that are finally saved to the database. The difference between edit requests and edit operations is that the first are issued when users just click on the ”edit“ tab placed on top of the articles’ pages whereas the latter are generated when users indicate a write operation to the database to save their changes or their contributed contents. Submit operations are those directed to preview the result of the modifications performed on the current content of an article or to highlight the differences introduced by a given edit operation in curse. History requests present the different revisions (edit operations) performed on an article’s content and leading to its actual version and state.

According to Figures 3 and 4, only those URLs involving visits, searches and edit requests would exhibit temporal repetitive patterns. On the other hand, requests consisting in edits (save operations), history reviews or submits for previewing contents would not present such cyclical evolutions over time. This is likely due to the fact that the requests exhibiting repetitive behaviors correspond to the most usual or generalized types of requests that compose the traffic to Wikipedia. The other kind

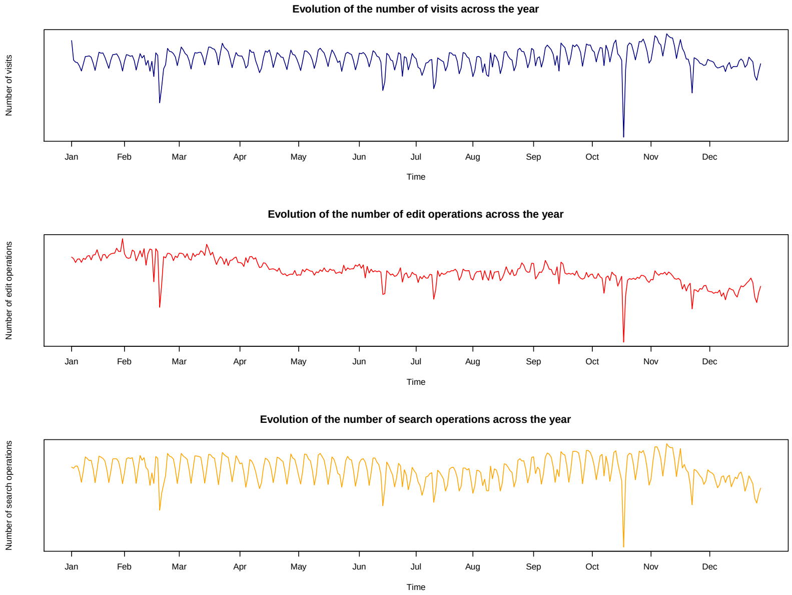

Fig. 4 Evolution of visits, edit operations and search requests throughout 2009.

of requests, on the contrary, have a more specialized character and because of this they appear rarely in the traffic. As a result, the most common requests follow the same periodical evolution than the general traffic to Wikipedia whereas the rest of requests show a more spurious behavior.

We undertake now the same analysis focusing on every whole week during 2009. The aim is to determine whether there are patterns involving any type of requests that are repeated along every week of the year. This is done, for example, in Figure 5 for the German, English, Spanish and French Wikipedias. This closer perspective confirms the similar weekly evolution of visits, searches and edit requests in contrast to the spurious and irregular nature of the requests consisting in edit operations, history and submits.

We decided to undertake the study of the evolution of visits and edits at the level of the days of the week in the aim of finding a meaningful closeness between their two temporal variations. As a result of such kind of analysis, Figure 6 presents the evolution of both types of requests throughout the days of the week for all the considered Wikipedias. Visits and edits, in each Wikipedia edition, correspond to the overall year and have been grouped by their day of issue. So, Figure 6 presents their compared progressions and shows a considerably closeness in the evolution of both types of requests in several Wikipedias. Nevertheless, the number of edits tends to raise in weekends for a group of them (French, Japanese, Dutch and Polish). That could mean that, in those editions, editors are not part of the great mass of people visiting the articles but just a minor group devoted to contribute or to maintain them.

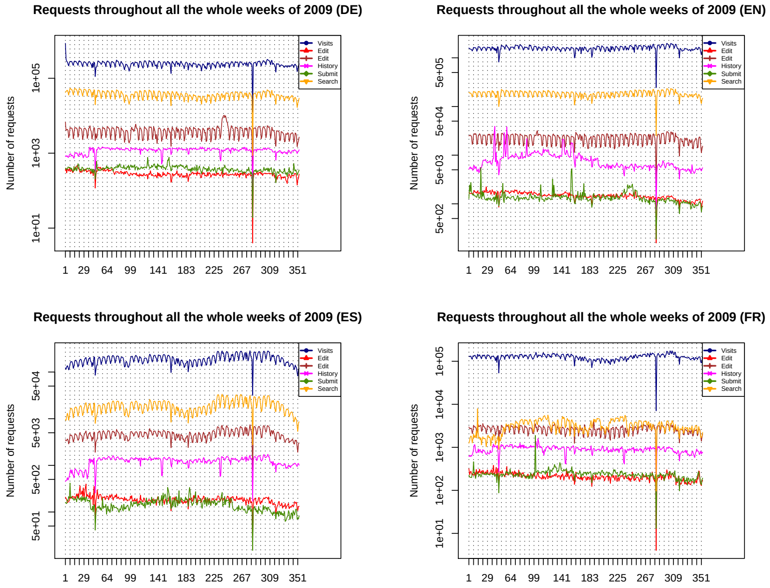

Fig. 5 Evolution of the different types of requests during every whole week of 2009 (DE EN ES FR).

5 Comparing the number and temporal evolution of the actions requested to Wikipedia

Figures 7 and 8 present the monthly evolution of edit requests, edit operations, history, submit and search requests for the considered Wikipedias. Although these figures are very similar in scale, we have preferred to present them using a logarithm scale in order to obtain more differentiated lines and, by means of this, a higher level of detail. As it can be observed from the chart, search operations are the most numerous actions in all the Wikipedias followed by the edit requests. As we can see, edit requests are considerably higher in number than edit operations. This means that an important number of edit requests are not finished by the corresponding write request to the database. Moreover, edit (write) operations are always very near the submit ones, which means that most of users regularly preview their changes before indicating their permanent storing to the database. In respect to the temporal evolution, edit requests and searches, again, present relatively similar evolutions as visits are not considered in this examination.

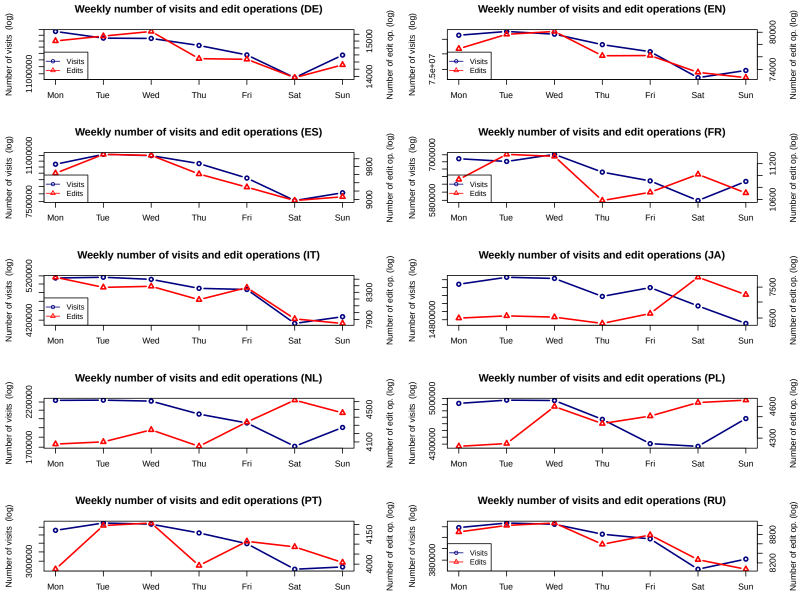

Fig. 6 Evolution of visits and edits throughout the days of the week in the different editions of Wikipedia.

6 Conclusions and further work

We can extract several conclusions after our efforts for characterizing temporarily the requests submitted to Wikipedia. First of all, we have shown how temporal information related to users' requests can be obtained from log lines stored by Wikimedia Foundation's Squid servers. Using this information we have modeled the variations over time of the different kind of requests submitted by users to Wikipedia. Our first finding was the fact that requests to Wikipedia temporarily model the overall traffic to all the Wikimedia Foundation projects. Of course it was what we were expecting, as Wikipedia is, by far, the most trending project maintained by the Wikimedia Foundation. However, we managed to obtain a high degree of correlation between Wikipedia's traffic and the requests directed to all the Wikimedia Foundation projects. In addition, we have illustrated how demands to Wikipedia consisting in visits, searches and edit requests present repeated patterns over time as they are the most generally solicited. On the other hand, submit or history requests and edits present a spurious and irregular nature because of their most specific character. In relation to this topic, the size of the sample may be determinant as the low percentage of edits contained in it can prevent the observation of cyclical distribution. So, further examinations should involve higher sampling factor to accurately analyze the presence of stationarity in the distribution over time of edits.

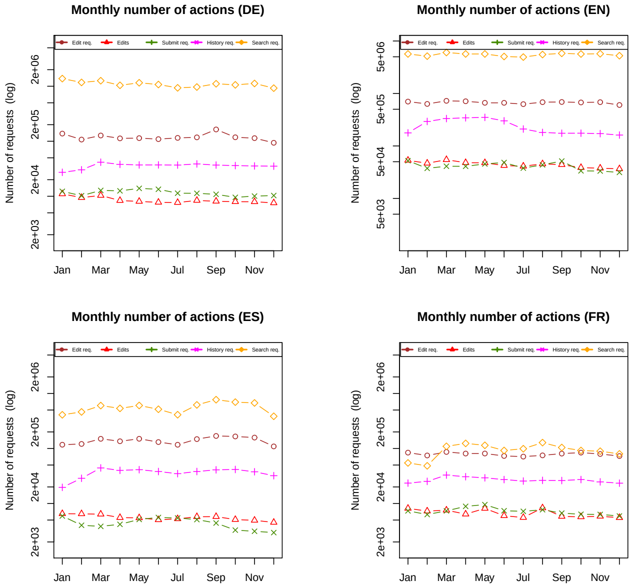

Fig. 7 Monthly distribution of the different types of actions in different Wikipedias.

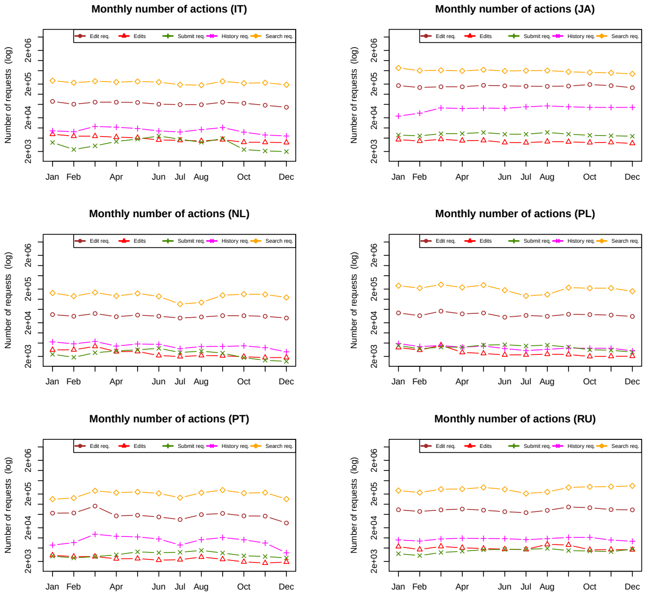

Fig. 8 Monthly distribution of the different types of actions in different Wikipedias.

Though the quantitative analysis of requests to Wikipedia may be considered tangential to the main aim of this paper, we consider that some observed findings in this area deserve to be mentioned. In this way, we have been able to appreciate how searches and edits are, respectively, the most and the least requested types of actions. Interestingly, we have shown how there is a significant relevance between the number of edit requests and the writes operation to the database that indicates that edit requests are abandoned by users in a considerably number of times. On the other hand, edits and submit requests remains very similar in number, which means that users usually exhibit the adequate habit of previewing changes before applying them to be permanent.

In the future, we plan to add geolocation to the temporal characterization process. In this way, a reference time plus the geographical position could better serve to determine the habits of the different communities of users when browsing Wikipedia. Furthermore, a closer analysis of the evolution of the different types of requests will allow to find more accurately defined relationships among them.

References

1. T. B. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to wikipedia content. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, 2008. ACM Press.
2. T. B. Adler, L. de Alfaro, I. Pye, and R. V. Measuring author contributions to the wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, New York, NY, USA, 2008. ACM Press.
3. A. Halavais and D. Lackaff. An analysis of topical coverage of wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440, 2008.
4. S. Javanmardi, Y. Ganjisaffar, C. Lopes, and P. Baldi. User contribution and trust in wikipedia. In *Collaborative Computing: Networking, Applications and Worksharing, 2009 CollaborateCom 2009. 5th International Conference on*, pages 1 –6, nov. 2009.
5. F. Olleros. Learning to trust the crowd: Some lessons from wikipedia. In *e-Technologies, 2008 International MCETECH Conference on*, pages 212 –216, jan. 2008.
6. F. Ortega, J. M. Gonzalez-Barahona, and G. Robles. 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, July 2007.
7. R. Priedhorsky, J. Chen, Shyong, K. Panciera, L. Terveen, and John. Creating, destroying, and restoring value in wikipedia. *MISSING*, November 2007.
8. B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The singularity is not near: slowing growth of wikipedia. In *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, pages 1–10, New York, NY, USA, 2009. ACM.
9. S. Tony and J. Riedl. 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, 2009. ACM.
10. G. Urdaneta, G. Pierre, and M. van Steen. A decentralized wiki enginge for collaborative wikipedia hosting. In *Proceedings of the 3rd International Conference on Web Information Systems and Technologies*, pages 156–163, March 2007.
11. D. M. Wilkinson and B. A. Huberman. Assessing the value of coooperation in wikipedia, April 2007.
12. J. Willinsky. What open access research can do for wikipedia. *First Monday*, 12(3), March 2007.